JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 4, pp. 889-899, Apr. 2026

Check for updates

의도적 결함 기반 UX를 통한 AI 사용성 평가 및 성능 분석

원 수 정[1] · 김 성 우[2*]
[1]한림대학교 디지털인문예술전공 학부생
[2]한림대학교 디지털인문예술전공 부교수

Assessing AI Usability and Performance Using Deliberately Flawed UX

Su Jeong Won[1] · Sung Woo Kim[2*]

[1]Undergraduate Student, Digital Arts & Humanities Major, Hallym University, Chuncheon 24252, Korea
[2]Associate Professor, Digital Arts & Humanities Major, Hallym University, Chuncheon 24252, Korea

[요 약]

이 연구의 목적은 AI의 사용성 평가 역량을 실증적으로 검증하고, 사용성 평가에서 AI의 적용 가능 범위와 인간 평가자와의 관계를 규명하는 데에 있다. 이를 위해 네 개의 AI 모델을 대상으로 두 가지 실험을 수행하였다. 첫 번째 실험에서는 사용성 결함을 의도적으로 포함한 UX를 제작하고 AI로 평가를 진행하여 문제 탐지 성능을 분석하였다. 두 번째 실험에서는 상용 앱을 대상으로 AI와 인간 전문가가 동일한 지침에 따라 평가를 수행하도록 하여 결과를 비교하였다. 또한 이미지와 동영상 입력 조건에 따른 AI의 평가 성능 차이도 분석하였다. 연구 결과, AI는 정적 UI 요소의 사용성 문제 식별에는 일정 수준의 성과를 보였으나, 화면 전환에 수반되는 인터랙션 흐름과 맥락적 사용성 판단에서는 한계를 드러냈다. 이는 AI가 인간 평가자를 대체하기보다는 보완적 도구로 활용되는 것이 적절하며, 향후 다양한 요인에 대한 통합적 해석 능력이 강화되어야 함을 시사한다.

[Abstract]

This study empirically investigates the usability-evaluation capabilities of artificial intelligence (AI) and its role in relation to human evaluators. Two experiments were conducted using four AI models. First, a user experience (UX) prototype containing usability flaws was developed and evaluated by AI to analyze issue-detection performance. Second, AI and human experts independently evaluated a commercial application, and their results were compared. Additionally, differences in AI performance under image and video-based input conditions were analyzed. The results show that AI was moderately effective when static UI elements were involved but exhibited limitations in interaction flows associated with screen transitions and context-dependent issues. These findings suggest its role as a complementary tool and highlight the necessity to enhance its ability to integrate diverse usability factors.

색인어 : 사용성 평가, 인공지능, UX, 사용성, 멀티모달 프롬프트
Keyword : Usability Evaluation, AI, UX, Usability, Multimodal Prompting

http://dx.doi.org/10.9728/dcs.2026.27.4.889

Received 30 December 2025; **Revised** 04 February 2026
Accepted 05 March 2026

***Corresponding Author; Sung Woo Kim**

Tel: +82-33-248-1515
E-mail: caerang@hallym.ac.kr

Ⅰ. 서 론

인공지능(AI), 특히 대규모 언어모델(LLM)의 발전은 사용자 경험 평가 방식에도 새로운 가능성을 제시하고 있다. 최근 생성형 AI는 다양한 영역에서 사용되며 사용자 인터페이스(UI)를 자동으로 생성하거나 설계를 지원하는 도구로까지 발전하고 있다. 이러한 흐름 속에서 LLM은 텍스트뿐 아니라 이미지와 동영상까지 포함한 멀티모달 입력을 처리할 수 있게 되면서 자동화된 UX 평가의 일부 과정을 대체하거나 보조할 수 있는 잠재적 도구로 주목받고 있다[1]. 이에 따라 인간 전문가 중심으로 수행되던 사용성 평가 역시 AI 기반 자동화 대상으로 확장되고 있다. 사용성 평가(Usability Evaluation)는 사용자의 실제 사용 행태(Use Behavior), 사용 맥락(Use Context), 기대된 심성 모델(Mental Model)[2], 과업 수행 과정 전반을 종합적으로 해석하면서 수행하는 활동으로 고도의 전문적 판단을 요구한다[3]. AI가 이러한 복합적인 사용성 요인을 통합적으로 분석하면서 인간 UX 전문가와 유사한 수준의 사용성 품질을 검사할 수 있는지에 대해서는 체계적인 실증 연구가 아직 제한적이다.

최근 AI 기반 자동 사용성 평가에 관한 연구가 증가하면서 멀티모달 LLM을 활용한 평가가 시도되고 있다. 그런데도 기존 연구는 정적 이미지 분석에 집중되어 실제 사용 맥락에서의 동적 상호작용과 사용자 심성 모델을 충분히 반영하지 못했다. 이에 따라 AI 기반 사용성 평가의 가능성을 검증하기 위해 추가적인 실증 연구가 요구된다.

이 연구는 이미지와 동영상을 포함한 멀티모달 입력 환경에서 모바일 앱 사용성 평가에 대한 AI의 정확성과 한계를 실증적으로 검증하고자 한다. 이를 위해 의도적으로 사용성 결함을 포함한 프로토타입을 활용하여 AI의 문제 탐지 능력을 평가하고, 실제 상용 앱 환경에서 인간 전문가와의 비교를 통해 AI의 활용 가능성을 규명하고자 한다.

이를 위해 두 단계의 실험을 수행하였다. 첫 번째 실험에서는 제이콥 닐슨의 10개의 휴리스틱스(Heuristics) 원칙[4], 애플 UX의 디자인 시스템인 휴먼 인터페이스 가이드라인(HIG)[5], 그리고 문화적 사용성(Cultural Usability)이라는 세 가지 기준을 기반으로 의도적으로 저급한 수준의 사용성을 탑재한 온보딩 프로토타입을 제작하여 AI가 기준을 위반하는 사용성 문제를 얼마나 정확하게 탐지하는지 평가하였다.

두 번째 실험에서는 실제 상용 앱을 대상으로 인간 전문가와 AI가 동일한 평가 지침에 따라 독립적으로 사용성 평가를 수행하도록 하여 양자의 문제 인식 수준 및 판단 경향의 유사성과 차이를 비교하였다. 이를 통해 AI가 사용성 평가에서 어느 수준까지 활용될 수 있는지, 인간 전문가의 역할을 어느 정도 보조하거나 대체할 수 있는지 탐구하였다.

이 연구에서는 다음의 연구 질문을 설정한다. AI는 사용성 품질이 떨어지는 UX 요소를 얼마나 정확하게 탐지할 수 있는가? (RQ1), 사용성 평가에서 AI와 인간 전문가의 판단은 어떤 차이를 보이는가? (RQ2), AI는 UX 평가에서 인간 전문가를 대체하거나 보조할 수 있는가? (RQ3)

Ⅱ. 선행 연구 고찰

2-1 AI 기반 자동 사용자 경험 평가의 진전

최근 자동화 사용성 평가에 AI를 적용하려는 연구가 증가하고 있다. Liu의 체계적 문헌 고찰에 따르면 머신러닝과 LLM, 생성형 AI는 사용자 행동 분석과 피드백 해석, 자동 휴리스틱 평가 과정에서 평가의 효율성과 확장성을 향상시킬 수 있다[6]. 그러나 AI 기반 사용성 평가는 판단 근거와 의사결정 과정이 명확히 설명되지 않는 특성을 보이며 환각, 맥락 오해, 데이터 편향으로 인해 실제 사용 상황을 충분히 반영하지 못할 수 있다는 한계가 지적된다.

최근 멀티모달 LLM은 텍스트·이미지 입력을 동시에 처리하며 더 정교한 의미 기반 분석을 수행할 수 있다는 점에서 주목받고 있다. 대표적으로 Lubos 외의 연구에서는 멀티모달 LLM이 정적 스크린숏을 기반으로 UI의 구조적·시각적 문제를 탐지하고, 그 결과를 인간 전문가의 휴리스틱 평가와 비교함으로써 LLM이 일정 수준의 품질 판단 능력을 보유함을 보여주었다[1]. 그러나 입력이 정적 화면에 한정되어 있어, 실제 사용 과정에서 발생하는 상호작용 흐름, 피드백 타이밍, 심성 모델 충돌과 같은 동적 사용성 문제를 포착하지 못했다는 한계가 있다.

2-2 시선 추적(Eye-Tracking) 기반 LLM 활용 연구

시선 추적 기술을 LLM 기반 사용성 평가와 결합하려는 시도도 최근 등장하고 있다. Kadegaonkar 외의 연구에서는 실시간 웹캠 기반 Eye-Tracking과 멀티에이전트 모델을 활용하여 사용자가 정적 UI 이미지를 주시하는 패턴을 분석하고, 이를 바탕으로 UI의 강점과 약점을 자동으로 진단하는 시스템을 제안하였다[7]. 이 연구는 특히 LLM이 시선 데이터의 주목도를 해석해 UI 구성 요소의 문제 가능성을 설명하고, 평가 결과를 구조화된 보고서 형태로 생성할 수 있음을 보여주었다는 점에서 의의가 있다. 그러나 해당 연구 역시 정적 UI 이미지로 제한된 입력으로 실제 동적 상호작용 흐름(전환, 스크롤, 상태 변화)이나 문화적·플랫폼 규범 기반의 오류를 평가하지는 못했다. 또한 시선 패턴은 UI의 시각적 복잡성이나 배치 문제를 탐지하는 데 유효하나 기능적 오류나 심성 모델 충돌 등을 직접적으로 감지하기 어렵다는 한계가 존재한다. 따라서 사용성 문제를 더욱 종합적으로 분석할 수 있는 평가 방법이 필요하다.

2-3 모바일 앱 휴리스틱 평가와 문화적 사용성

모바일 앱 사용성 평가는 UI 규칙 위반 탐지를 넘어 사용 환경과 문화적 맥락을 반영한 휴리스틱 적용의 필요성이 제기되어 왔다. 문화적 요소를 고려한 평가는 단순히 문제의 양

을 늘리는 것이 아니라 사용자 경험에 실질적 영향을 미치는 핵심 문제를 효과적으로 식별한다[8]. 이에 따라 모바일 특화 휴리스틱을 한국 문화권에 맞게 재구성하고 현지화하여 전문가 검증을 통해 언어적·문화적 적합성과 평가 해석의 일관성을 확보한 평가 도구가 제안되었다[9]. 이는 문화적 타당성이 모바일 사용성 평가 결과에 중대한 영향을 미침을 시사한다.

Ⅲ. 연구 방법

3-1 실험 1. 저품질 사용성 UX 요소의 탐지 성능

첫 번째 실험에서는 AI가 저급의 UX 요소를 탐지하는 성능을 탐구하였다.

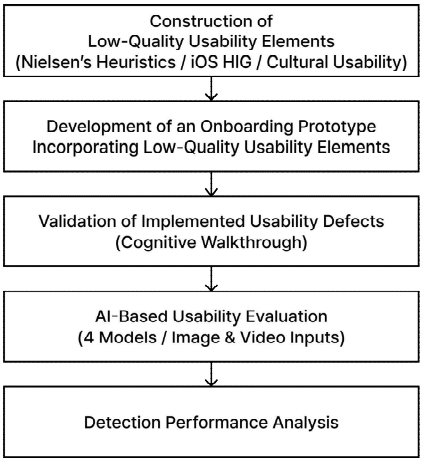


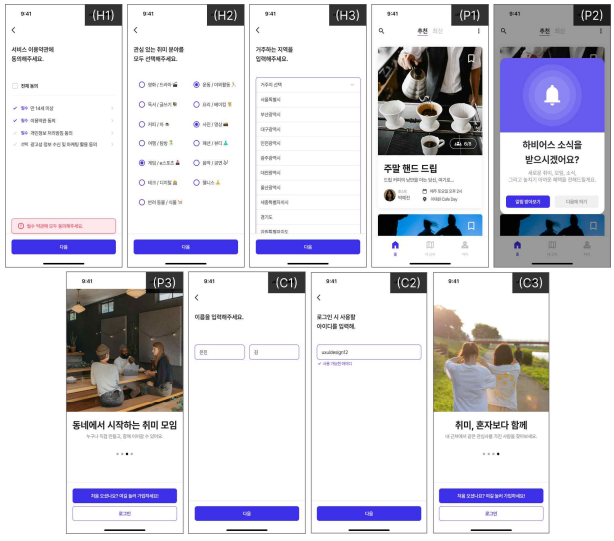
그림 1. 저품질 사용성 UX 요소의 탐지 성능 평가 방법
Fig. 1. Evaluation procedure for detecting low-quality usability elements

사용성 평가 기준으로는 제이콥 닐슨의 10가지 휴리스틱, 플랫폼 디자인 시스템(iOS HIG), 문화적 사용성의 세 축을 적용하였다. 휴리스틱은 보편적 사용성 원칙을 제공하며 디자인 시스템은 운영체제 고유의 상호작용 규범과 일관성을 보장한다[10]. 또한 문화적 사용성은 언어·상징·정보 구조 등 문화권 별 사용자 특성을 반영하여 실제 사용 환경에서 발생할 수 있는 문제를 식별하는 데 필수적이다[11]. 이 세 가지 기준을 기반으로 저급한 품질의 UX 요소를 고안하였다.

표 1의 요소들은 보편적 사용성 원칙, 플랫폼의 네이티브 규범, 그리고 문화적 맥락이라는 서로 다른 차원에서 발생하는 사용성 위반을 유형화한 것이다. 이 요소들을 적용하여 먼저 그림 1처럼 저급한 사용성 품질로 되어 있는 모바일 앱 내 온보딩(Onboarding) 부분의 프로토타입을 제작하고, 연구팀 소속 UX 전문가급 1인 및 주니어급 3인이 사전에 Cognitive Walkthrough를 수행하여 사용성이 좋지 않은 UX 요소들이 자연스럽게 탑재되어 있는지 검증하였다.

표 1. 저품질 사용성의 UX 요소
Table 1. Low-quality usability UX elements

Criteria	ID	Element
Jakob Nielsen's 10 Usability Heuristics	(H1)	Next button enabled without required selection
	(H2)	Mismatch between using radio buttons in a multiple-selection context
	(H3)	Lack of feedback where icon state changes do not reflect activation
Platform Design System (iOS HIG)	(P1)	Visual inconsistency due to non-uniform icon line thickness and details
	(P2)	Button placement in a two-button dialog that conflicts with cognitive flow
	(P3)	Use of emotional or abstract labels that hinder direct meaning delivery
Cultural Usability (Korea)	(C1)	Western-style input order for name fields (First name–Last name)
	(C2)	Left–right screen transition direction inconsistency
	(C3)	Mixing honorific and non-honorific language



*Written in Korean to evaluate cultural usability in the Korean context.
그림 2. 저품질 사용성의 UX 요소가 반영된 모바일 앱 온보딩 프로토타입(일부)
Fig. 2. A mobile app onboarding prototype incorporating low-quality usability UX elements (excerpt)

이 연구에서는 기존 데이터 이력을 초기화한 상태의 ChatGPT, Gemini, Claude, UX Pilot의 네 가지 AI 모델에게 사용성 평가를 수행하게 하였다. 표 2와 같이 ChatGPT, Gemini, Claude의 경우 화면 스크린숏 이미지와 사용 과정을 녹화한 동영상으로 평가를 각각 수행하였으며, 평가 목적과 기준 그

리고 각 평가가 어느 화면에서 도출되었는지를 명시하도록 요구하는 동일한 프롬프트를 부여하였다. 반면 UX Pilot[12]은 동영상 입력은 안 되고 이미지만 입력되는 제약이 있어 화면 스크린숏 형태로만 정보를 제공하여 평가를 수행하게 하였으며, 평가 기준과 목적만 입력할 수 있는 도구의 특성을 고려하여 해당 항목만 제시하였다. 아울러 본 연구의 목적은 각 AI 모델 간 성능의 우열을 판단하는 데 있지 않다.

표 2. AI 모델별 입력 정보의 미디어 유형
Table 2. Input media types across AI models

	Image	Video
ChatGPT	○	○
Gemini	○	○
Claude	○	○
UX Pilot	○	-

3-2 실험 2. 인간 전문가와 AI의 사용성 평가 비교 분석

두 번째 실험에서는 AI가 실제 상용 앱의 사용성 문제를 인간 UX 전문가에 비해 얼마나 식별해 내는지를 파악하고자 하였다.

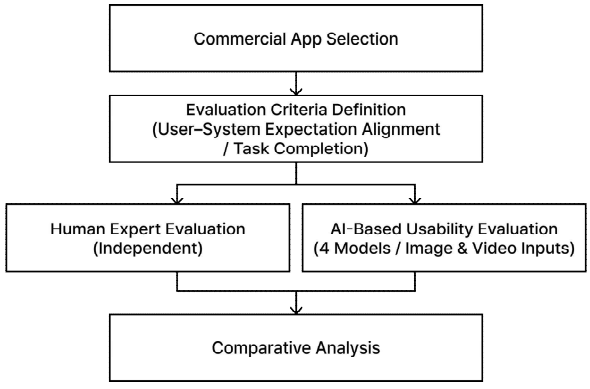


그림 3. 인간 전문가와 AI의 사용성 평가 비교 분석 방법
Fig. 3. Experimental procedure for AI-human expert comparative usability evaluation

대상 앱은 음성 녹음, 문항 풀이, 결과 확인 등 다양한 상호작용이 포함된 'Y사의 토익 기출문제 프리뷰' 앱으로 선정하였다. 사용성 평가는 사용자 기대와 시스템 동작의 일치 여부[13], 그리고 과업 완료 가능성[14]의 두 기준에 따라 이루어졌다. 이를 위해 UX 전문가급 1인 및 주니어급 3인이 사전 Cognitive Walkthrough를 통해 앱의 주요 사용성 문제점을 검토하였다. 이후 경력 5년 이상의 외부 UX 전문가 2인이 독립적으로 평가를 수행하였다. 전문가들이 공통으로 식별한 사용성 문제를 AI 평가의 주요 기준점으로 설정하였으며, 이는 합의 기반 정답 집합(consensus-based ground truth) 구성 방식에 해당한다. AI 평가는 앞선 실험과 동일하게 기존 데이터 이력을 초기화한 후 표 2 조건에 따라 수행되었으며, AI의 평가 결과는 인간 전문가의 평가를 기준으로 다음 세 가지 범주로 분석하였다. 1) 전문가와 AI가 동일하게 문제로 식별한 항목, 2) 전문가는 문제로 보았으나 AI가 탐지하지 못한 항목, 3) 전문가는 문제로 보지 않았음에도 AI가 문제로 제기한 항목이 이에 해당한다.

이용자 흐름(User Flow)은 사용자가 특정한 의도와 기대를 하고 앱을 조작하는 과정에서 UI 상태와 인터랙션이 순차적으로 전개되어 하나의 의미 있는 과업을 완성하는 흐름을 의미한다[15]. 또한 인터랙션 흐름(Interaction Flow)은 사용자가 특정 목표를 달성하기 위해 인터페이스와 상호작용을 하는 과정에서 경험하는 행동의 순서와 상태 전이의 흐름을 의미한다[16]. 즉 UI 사용 과정에서 행동-반응-상태 변화가 연결되면서 누적되는 동적 구조에 해당한다. 이 연구에서는 사용자 흐름에 따라 두 개 이상의 화면의 전환과 그 과정에서 누적되는 상호작용의 동적인 구조를 인터랙션 흐름으로 규정하고 실험을 진행하였다. 실험 2에선 인터랙션 흐름에 대해 인간 전문가가 지적한 사용성 문제를 AI가 어느 정도 탐지할 수 있는지도 분석한다.

Ⅳ. 연구 결과

4-1 실험 1 결과

AI 모델이 입력받는 미디어 타입에 따라 탐지 능력에서 뚜렷한 차이를 보인다는 점을 확인하였다. 표 3~표 6에서 ○는 해당 위반 요소를 AI가 탐지했음을, ×는 탐지하지 못했음을, △는 AI가 언급했으나 문제가 아니라 긍정적 요소로 판단한 경우를 의미한다.

표 3. 이미지 미디어 입력에 따른 AI 모델별 저품질 UX 요소 탐지 결과
Table 3. Low-quality UX detection across AIs (image input)

	H1	H2	H3	P1	P2	P3	C1	C2	C3
ChatGPT	○	○	×	×	×	△	○	×	△
Gemini	○	○	×	×	×	○	○	×	○
Claude	○	○	×	×	×	△	×	×	×
UX Pilot	○	×	×	×	×	×	×	×	×

ChatGPT는 화면 스크린숏 조건에서 (H1), (H2), (C1)을 정확하게 탐지하였다. 그러나 (P3)와 (C3)는 명확한 위반 요소임에도 이를 긍정적 요소로 해석하며 문제로 간주하지 않았다. (P3)는 '직관적 의미 전달을 방해하는 감성적·추상적 레이블 사용'으로 분류되는 HIG 위반 요소임에도 불구하고 ChatGPT는 문제로 판단하지 않았다. 오히려 '구어체·친근체 표현이 MZ 세대의 선호와 부합한다'라는 이유로 해당 레이블

이 한국 문화와 조화를 이룬다고 긍정적으로 평가하였다. 마찬가지로 (C3) 역시 '존댓말과 반말의 혼용'이라는 명확한 한국어 문화 관습 위반임에도 이를 '한국 MZ 세대 및 모바일 앱 문체 트렌드'로 해석하여 문제로 판단하지 않았다. 이는 ChatGPT가 언어 사용 규범과 같은 문화적 사용성 판단이 요구되는 요소에서는 일부 사용자 집단의 특성을 전체 사용자의 사용 행태로 일반화하여 평가 근거로 설정하는 한계를 보였다.

표 4. 이미지 미디어 입력에 따른 AI 모델별 저품질 UX 요소 탐지 비율

Table 4. Distribution of detection outcomes across AIs (image input)

	○ (Detected)	× (Missed)	△ (Misjudgment)
ChatGPT	33.3%	44.4%	22.2%
Gemini	55.6%	44.4%	0%
Claude	22.2%	66.7%	11.1%
UX Pilot	11.1%	88.9%	0%

표 5. 동영상 미디어 입력에 따른 LLM별 저품질 UX 요소 탐지 결과

Table 5. Low-quality UX detection across LLMs (video input)

	H1	H2	H3	P1	P2	P3	C1	C2	C3
ChatGPT	×	×	×	×	×	○	×	×	×
Gemini	○	×	×	×	×	×	○	×	×
Claude	×	×	×	×	×	○	○	×	×

Gemini는 화면 스크린숏 조건에서 (H1), (H2), (P3), (C1), (C3)를 정확하게 탐지하였다. 특히 ChatGPT가 문제로 분류하지 않았던 (P3)와 (C3)를 명확히 위반 요소로 식별했다는 점에서 Gemini는 정적 화면을 기반으로 한 플랫폼 규범 위반 요소와 문화적 관습 위반 요소를 비교적 넓은 범위에서 포착하였다. 그럼에도 Gemini 역시 (H3), (P1), (P2), (C2) 항목들은 탐지하지 못했다. 이는 Gemini 또한 정적 정보 기반 판단에는 강점을 보이지만, 사용자의 조작에 따른 상태 변화나 플랫폼 규범의 세부적인 구조와 같은 UX 원칙을 처리하는 데에는 한계가 있음을 의미한다.

Claude는 화면 스크린숏 조건에서 (H1)와 (H2)를 정확하게 탐지하였다. 그러나 (P3)에 대해서는 '직관적 의미 전달을 방해하는 감성적·추상적 레이블 사용'임에도 불구하고 오히려 '주요 액션을 유도한다'는 이유로 긍정적으로 평가하였다. 이는 Claude가 사용자의 인지적 이해 가능성보다는 행동 유도 여부와 기능적 효과를 중심으로 사용성 품질을 판단했음을 보여준다. 그 결과, 의미 전달을 저해하는 사용성 위반 요소조차 목적 달성에 기여한다는 이유로 긍정적으로 평가하였다.

UX Pilot은 단일 화면 스크린숏 분석 기능을 중심으로 작동하는 도구의 특성상 탐지 범위가 가장 제한적이었다. 이 실험에서 UX Pilot이 정확히 탐지한 요소는 (H1) 단일 항목이었으며 모든 나머지 요소는 탐지하지 못했다. 이는 고정된 UI 구조 중심의 해석 방식이 시각적·의미적·문화적 위반이 복합적으로 얽혀 있는 UX 문제를 포괄적으로 파악하기에는 근본적 제약을 지니고 있음을 보여준다.

표 6. 동영상 미디어 입력에 따른 AI 모델별 저품질 UX 요소 탐지 비율

Table 6. Distribution of detection outcomes across AIs (video input)

	○ (Detected)	× (Missed)
ChatGPT	11.1%	88.9%
Gemini	22.2%	77.8%
Claude	22.2%	77.8%

동영상 기반 조건에서 세 가지 LLM 모델은 의도적으로 탑재한 사용성 문제점 대부분을 탐지하지 못했으며, ChatGPT는 (P3), Gemini는 (H1)과 (C1), Claude는 (P3)와 (C1)만을 식별하는 데 그쳤다. 이는 사용자 조작에 따른 상태 변화나 화면 전개 과정에서의 표현 일관성과 같이 동영상으로부터 제공되는 추가적인 정보를 LLM이 보편적 사용성 원칙, 플랫폼의 네이티브 규범, 문화적 맥락에 따른 사용성 품질 파악에 연결 짓지 못하고 있음을 보여준다. 즉, 이미지 대비 더 많은 정보를 동영상이 제공함에도 불구하고 AI가 이를 사용성 평가에 효과적으로 활용하지 못한다는 것이다.

실험 1의 결과를 요약하면 다음과 같다. 설계된 9개의 위반 요소 중 ChatGPT는 동영상 조건에서 1개 및 화면 스크린숏 조건에서 3개를 탐지하였으며, Gemini는 각각 2개와 5개, Claude에서는 각각 2개씩 탐지하였다. UX Pilot은 1개만 탐지하여 4개 모델 중 탐지율이 가장 낮았다. 또한 현 단계의 LLM은 화면 스크린숏 입력 조건에서는 UI 구성상 명확하게 드러나는 저품질 사용성 요소를 비교적 안정적으로 탐지하는 반면에 사용자의 사용 흐름과 상호작용 과정에 대한 더 풍부한 정보를 제공하는 동영상 입력 환경에서 이미지 입력 상황 이상의 탐지 성능을 발휘하지 못하였다. 즉, 동영상이라는 정보 측면으로 확장된 입력이 사용성 평가 성능의 질적 향상으로까지 이어지지는 못하는 한계가 확인되었다.

4-2 실험 2 결과

인간 평가자 및 AI 평가자가 수행한 결과를 (E1) 전문가와 AI가 동일하게 문제로 식별한 항목, (E2) 전문가가 문제로 보았으나 AI가 탐지하지 못한 항목, (E3) 전문가가 문제로 보지 않았음에도 AI가 문제로 제기한 항목의 세 가지 기준으로 정리하였다. 표 7과 표 9의 (E3)에서 ○는 AI의 사용성 평가가 타당한 경우를, ×는 타당하지 않거나 과도한 평가로 판단된 경우를 의미한다. △는 사용성 이슈를 지적하였으나 근거가 명확하지 않은 경우를 나타낸다. (E3)에 대한 AI 사용성 평가의 적절성 판단은 연구팀 소속 UX 전문가 1인과 주니어 평가

표 7. 인간 전문가와 AI의 사용성 평가 결과(이미지)
Table 7. Evaluation results by human experts and AI (image input)

	Screenshot-based usability evaluation	
(E1)	The “CLOSE” button is placed in the top action area rather than the center of the screen, making it difficult to locate intuitively. (Gemini, Claude, UX Pilot)	
(E2)	The app launches in landscape orientation by default.	
	“Register for Test” and “SOUND TEST” have different functions and risk levels but are placed side by side, increasing mis-tap risk.	
	The problem-type selection appears like a list title and is not recognized as a button.	
	Users expect the test to start after pressing “START”, but problem-type selection interrupts the flow.	
	The “Register for Test” function leads users out of the app, breaking the task flow.	
	Selected options cannot be deselected, conflicting with common select-cancel interaction models.	
	Although “Direction 1 of 3” is displayed, there is no visual progress indicator.	
	Voice guidance similar to actual TOEIC speaking tests increases immersion (positive element).	
	Guidance relies on text and tables, reducing readability and comprehension under pressure.	
	Preparation and response times share the same UI, making the states hard to distinguish.	
	Instruction text in the automatically opened popup at the end of a response is too long, and the popup closes too quickly to be fully understood.	
	The purpose of the Direction screen is unclear, despite expectations that “START” would begin the test.	
	The absence of sound cues makes it difficult to recognize when speaking has ended.	
	The flow is inefficient because users cannot directly navigate between Response Check screens for different questions.	
	The playback bar is positioned at the top of the screen, reducing accessibility.	
	It is difficult to identify which question the current response screen corresponds to.	
	Only the question number is shown, requiring extra interaction to identify the question.	
	The replay button always restarts from the beginning, conflicting with users’ mental models.	
	Users can enter the Response Check screen without completing the test, resulting in an inefficient flow.	
	A separate stop button is provided instead of toggling play/stop, causing confusion.	
	Inconsistent terminology is used: other screens use “CLOSE”, while this screen uses "FINISH".	
	Numbered buttons suggest a fixed order of operation, reducing user autonomy.	
(E3)	There is a lack of language consistency between buttons and instructional text. (UX Pilot)	×
	It is difficult to tell whether scrolling is possible or how many test items remain. (ChatGPT, Claude)	×
	The placement of the NEXT button in the upper-right corner weakens its role as a CTA and disrupts visual flow. (ChatGPT)	△
	Despite being a popup, the absence of background dimming makes the buttons on the underlying screen appear active. (ChatGPT)	○
	The abbreviation “Que.” has very low intuitiveness. (ChatGPT, UX Pilot)	○
	It is difficult to determine whether the STOP button pauses everything or only an individual item. (ChatGPT, Claude)	△
	Outdated button styling, low text contrast, and skeuomorphic design weaken the distinction between active and inactive states. (ChatGPT)	○

표 8. 인간 전문가와 AI의 사용성 평가 결과 항목 분포 (이미지)
Table 8. Distribution of evaluation results by human experts and AI (image input)

Human Experts Comments (n = 25)		AI (Models: ChatGPT, Gemini, Claude, UX Pilot)
(E1) Expert-AI Agreement		1
(E2) Issues Missed by AI		22
(E3) Additional Issues Identified by AI	○ (Valid Issue)	3
	× (Overstated Issue)	2
	△ (Weakly Supported Issue)	2

자 1인이 공동으로 수행하였다. 다수 합의 기준에 따라 적절성을 검토하였으며 타당한 지적(○), 과도한 문제 제기(×), 사용성 문제 제기는 타당하나 근거가 부족한 평가(△)로 분류하였다. 표 8과 표 10은 인간 전문가와 AI 간 판단 일치도를 더욱 구체적으로 파악하기 위해 각각 표 5와 표 7의 평가 결과를 항목별 분포로 재구성한 것이다.

먼저, 전문가와 AI가 동일하게 사용성 문제로 식별한 항목(E1)은 1개로, 화면 상단 액션 영역에 배치된 “CLOSE” 버튼

표 9. 인간 전문가와 AI의 사용성 평가 결과 (동영상)
Table 9. Evaluation results by human experts and AI (video input)

	Video-based usability evaluation	
(E1)	The “CLOSE” button is placed in the top action area rather than the center of the screen, making it difficult to locate intuitively. (Gemini)	
(E2)	The app launches in landscape orientation by default.	
	The "Register for Test" button appears as a primary action, disrupting task prioritization.	
	“Register for Test” and “SOUND TEST” have different functions and risk levels but are placed side by side, increasing mis-tap risk.	
	The problem-type selection appears like a list title and is not recognized as a button.	
	Users expect the test to start after pressing “START”, but problem-type selection interrupts the flow.	
	The "Register for Test" function leads users out of the app, breaking the task flow.	
	Selected options cannot be deselected, conflicting with common select-cancel interaction models.	
	Voice guidance similar to actual TOEIC speaking tests increases immersion (positive element).	
	Instruction text in the automatically opened popup at the end of a response is too long, and the popup closes too quickly to be fully understood.	
	The purpose of the Direction screen is unclear, despite expectations that “START” would begin the test.	
	The absence of sound cues makes it difficult to recognize when speaking has ended.	
	The playback bar is positioned at the top of the screen, reducing accessibility.	
	It is difficult to identify which question the current response screen corresponds to.	
	The replay button always restarts from the beginning, conflicting with users’ mental models.	
	Users can enter the Response Check screen without completing the test, resulting in an inefficient flow.	
	A separate stop button is provided instead of toggling play/stop, causing confusion.	
	Inconsistent terminology is used: other screens use "CLOSE", while this screen uses "FINISH".	
(E3)	Text size and layout are not optimized for the mobile environment (Gemini).	○
	It is difficult to tell whether scrolling is possible or how many test items remain (ChatGPT).	×
	The BACK button is displayed in gray, but its active/inactive state is unclear (Claude).	×
	The “CLOSE” button lacks a confirmation dialog, increasing the risk of accidental termination (Claude).	○
	Although the numbers are only identifiers, they may be mistaken for key inputs, so the notation should be revised.	△
	Outdated button styling, low text contrast, and skeuomorphic design weaken the distinction between active and inactive states (Gemini).	○
	The absence of a waveform or volume meter makes it difficult to visually confirm whether voice recording is functioning properly, increasing user uncertainty (Gemini).	○

의 위치에 관한 문제였다. 해당 항목은 Gemini, Claude, UX Pilot이 공통으로 지적한 사항으로, 화면 스크린숏 기반 평가에서도 버튼의 위치와 시각적 위계와 같은 명시적인 레이아웃 문제는 AI가 전문가와 유사한 수준으로 인식한다는 것을 보여준다. 다만 (E1)에 포함된 항목 수가 매우 제한적이라는 점은 정적 화면만으로는 인간 전문가의 평가에 근접하는 사용성 평가를 수행하는 데 한계가 있음을 시사한다.

반면, 전문가가 문제로 판단하였으나 AI가 탐지하지 못한 항목(E2)은 22개로 다수 확인되었다. 여기에는 시각적 위계 부족, 심성 모델 불일치, 피드백 결핍, 일관성 문제로 인해 사용자의 과업 수행을 방해하는 사용성 문제들이 포함되며 현실과의 유사성을 통해 몰입감을 강화하는 긍정적 요소가 함께 나타났다.

전문가가 문제로 보지 않았으나 AI가 문제를 제기한 항목(E3)에서는 평가의 적절성이 혼재되어 나타났다. 팝업 화면임에도 흐린 배경(Deemed Screen) 처리가 없어 하위 화면의 버튼이 활성화된 것처럼 보이는 문제, 축약어의 직관성 부족, 버튼 스타일과 대비로 인해 활성·비활성 상태 구분이 어려운 점 등은 실제 사용성 관점에서도 타당한 지적(◯)으로 평가되었다. 반면, 스크롤 가능 여부나 남은 문항 수 인지에 대한 지적은 실제 화면 맥락을 충분히 반영하지 못한 과도한 평가(×)로 분류되었다. 또한 일부 지적(△)은 문제 제기 자체는 타당하지만 AI가 제시한 해석이 실제 사용 맥락과 완전히 일치하지는 않는 것으로 판단되었다. 예를 들어 다음(NEXT) 버튼의 상단 우측 배치는 CTA로서의 주목도가 낮다는 점에서 개선의 여지는 있으나, 이를 시선 흐름을 교란하는 문제로 해석할 만큼의 명확한 근거는 확인되지 않았다. 마찬가지로 정지(STOP) 버튼 역시 사용성 측면에서 혼란을 유발할 수 있는 요소는 존재하나, 전체 동작과 개별 항목을 구분하기 어렵다는 지적은 실제 인터랙션 설계의 핵심 문제와는 다소 차이가 있는 해석으로 평가되었다.

화면 스크린숏 기반 사용성 평가에서 AI는 시각적 대비, 버튼 스타일, 레이아웃과 같은 정적 UI 요소에 대해서는 일정 수준의 탐지 능력을 보였으나, 비효율적인 플로우, 불필요한 앱 이탈, 사용자의 기대와 상충하는 인터랙션 전개와 같은 인터랙션 흐름상의 사용성 문제는 충분히 판단하지 못했다. 이

표 10. 인간 전문가와 AI의 사용성 평가 결과 항목 분포(Video)
Table 10. Distribution of evaluation results by human experts and AI (Video Input)

Human Experts Comments (n = 25)		AI (Models: ChatGPT, Gemini, Claude)
(E1) Expert–AI Agreement		1
(E2) Issues Missed by AI		17
(E3) Additional Issues Identified by AI	◯ (Valid Issue)	4
	× (Overstated Issue)	2
	△ (Weakly Supported Issue)	1

는 화면 스크린숏이라는 제한된 입력 조건 하에서 AI 기반 사용성 평가의 적용 가능성과 한계를 동시에 보여준다.

마찬가지로 동영상 기반 평가 결과 역시 앞서 제시한 세 가지 기준 (E1), (E2), (E3)에 따라 정리하였으며, 전문가 판단과 불일치한 AI 평가(E3)에 대해서는 동일한 기준으로 적절성을 검토하였다.

먼저, 전문가와 AI가 동일하게 사용성 문제로 식별한 항목(E1)은 1개로 비교적 제한적으로 나타났다. Gemini는 화면 상단 액션 영역에 배치된 “CLOSE” 버튼의 위치가 직관적이지 않다는 문제를 전문가와 동일하게 지적하였다. 이는 동영상 기반 평가에서 AI가 화면 전환과 조작 흐름을 통해 사용자의 시선 이동과 조작 부담을 일정 수준 추론할 수 있음을 시사한다. 다만 (E1)에 포함된 항목 수가 많지 않다는 점은 실제 사용 맥락이 포함되었음에도 AI가 인간 전문가 수준으로 사용성 문제를 인지하는 데에 여전히 한계가 있음을 보여준다.

반면, 전문가가 문제로 판단하였으나 AI가 탐지하지 못한 항목(E2)은 17개로 다수 확인되었다. 여기에는 화면 스크린숏 기반 평가와 마찬가지로 시각적 위계 부족, 심성 모델 불일치, 피드백 결핍, 일관성 문제로 인해 사용자의 과업 수행을 방해하는 사용성 문제들이 포함되며 현실과의 유사성을 통해 몰입감을 강화하는 긍정적 요소가 함께 나타났다. 그러나 화면 스크린숏 기반 평가에서 전문가가 문제로 판단했으나 AI가 탐지하지 못한 항목(E2) 중 진행 단계에 대한 시각적 표시 부재, 텍스트·표 중심 안내로 인한 가독성 저하, Preparation/Response 단계 구분 불명확, 시험 문제 간 Response Check 화면 직접 이동 불가, 응답 화면에서 문제 식별 정보 부족, 버튼 부여로 인한 조작 자율성 제한의 6개 항목은 동영상 기반 평가에서는 정확히 탐지되었다. 이러한 결과는 동영상 기반 입력이 전반적인 사용성 문제 탐지 성능을 크게 향상하지는 못했으나, 화면 스크린숏 기반 평가에 비해서는 일부 사용성 문제를 추가로 식별할 수 있었음을 시사한다. 정적 화면만을 입력으로 제공한 경우에는 탐지되지 않았던 일부 항목이 동영상 조건에서는 탐지되었으며 이는 동영상 기반 평가가 화면 스크린숏 기반 평가보다 상대적으로 더 많은 사용성 정보를 제공할 수 있음을 보여준다.

전문가가 문제로 보지 않았으나 AI가 문제로 제기한 항목(E3)에서는 AI 평가의 적절성에 차이가 나타났다. 일부 항목의 경우(◯) 텍스트 크기 및 레이아웃 최적화 부족, 확인 다이얼로그에서의 종료 위험, 버튼 스타일 및 대비 문제, 음성 녹음 상태를 시각적으로 확인하기 어려운 점 등은 실제 사용성 관점에서도 타당한 지적으로 평가되었다. 이는 AI가 시각적 명확성이나 상태 인지와 관련된 문제를 비교적 일관되게 감지할 수 있음을 보여준다. 반면, 스크롤 가능 여부나 남은 문항 수 인지 문제(×)와 같이 실제 사용자 경험과는 다소 거리가 있는 항목도 제기되었다. 일부 항목(△)은 사용성 문제 제기는 타당했으나 명확한 근거는 부족한 것으로 판단되었다. 화면의 ‘[1]’ 숫자 레이블은 순서 표시 용도지만 조작 요소로 오해되었다. PC와 달리 스마트폰 환경에서는 키보드 숫자 단축키 입력이 불가능하고 입력을 유도하는 단서도 없어 실제로 조작해야 할 요소로 인식될 가능성은 낮다. 다만 버튼에 번호가 부여되어 순차적 조작을 강제하는 제약처럼 보일 수 있어 사용자 자율성을 저해하는 사용성 문제는 존재한다.

동영상 기반 사용성 평가에서 AI는 화면 스크린숏 기반 평가에 비해 일부 사용성 문제를 추가로 탐지하는 성과를 보였다. 특히 화면 전개 과정이나 상태 변화가 수반되는 항목 중 일부는 정적 화면만으로는 식별되지 않았으나 동영상 입력 조건에서는 탐지되었다. 다만 동영상 입력을 통해 AI의 평가 수행력이 다소 향상되기는 하나 전반적인 탐지 범위는 제한적이었으며 AI는 여전히 전문가가 지적한 다수의 사용성 문제를 탐지하지 못하였다. 이는 AI 사용성 평가에서 동영상 입력이 화면 스크린숏보다 더 많은 사용성 문제를 탐지할 수는 있지만 전문가 수준으로 사용성 문제를 포괄적으로 인식하는 데에는 아직 한계가 있음을 보여준다.

실험 2에서 인간 전문가가 지적한 인터랙션 흐름과 밀접한 사용성 문제는 6개로 (F1) 테스트 시작 이후 문제 유형 선택으로 인한 흐름 단절, (F2) 시험 접수 기능 버튼으로 인한 앱 이탈, (F3) 과도하게 길고 빠르게 종료되는 팝업 안내 문구, (F4) 시험 시작 직후 문제 풀이가 이루어질 거라는 기대를 저해하는 안내 화면 노출, (F5) 문제 간 답안 화면을 직접 이동할 수 없는 비효율적인 플로우, (F6) 시험 후 답안 화면으로 직접 이동해야 하는 번거로움이 이에 해당한다. 표 7에서 ◯는 AI의 각 미디어 입력 조건에서 해당 인터랙션 흐름 사용성 문제를 탐지했음을, ×는 탐지하지 못했음을 의미한다.

표 11과 같이 화면 스크린숏 기반 평가에서 AI는 인간 전문가가 지적한 인터랙션 흐름과 밀접한 사용성 문제 6개 전부 탐지하지 못했다. 이러한 문제들은 사용자의 조작 의도와 기대가 앱의 인터랙션과 일치하는지를 종합적으로 고려해야 판단할 수 있는 항목으로 화면 스크린숏 기반 AI 평가에서는 이

표 11. 인간 전문가와 AI의 인터랙션 흐름 사용성 평가 결과
Table 11. Interaction flow evaluation results by human experts and AI

	Interaction flow usability evaluation	Image	Video
(F1)	Users expect the test to start after pressing “START”, but problem-type selection interrupts the flow.	×	×
(F2)	The “Register for Test” function leads users out of the app, breaking the task flow.	×	×
(F3)	Instruction text in the automatically opened popup at the end of a response is too long, and the popup closes too quickly to be fully understood.	×	×
(F4)	The purpose of the Direction screen is unclear, despite expectations that “START” would begin the test.	×	×
(F5)	The flow is inefficient because users cannot directly navigate between Response Check screens for different questions.	×	○
(F6)	Users can enter the Response Check screen without completing the test, resulting in an inefficient flow.	×	×
–	The “CLOSE” button lacks a confirmation dialog, increasing the risk of accidental termination.	×	○

러한 인터랙션 흐름 특성을 충분히 반영되지 못한 것으로 나타났다.

반면 동영상 기반 평가에서 AI는 (F5)를 탐지했을 뿐만 아니라 인간 전문가가 탐지하지 못한 인터랙션 흐름 사용성 문제를 추가로 탐지하였다. 해당 문제는 사용자가 문제 풀이 중 닫기(CLOSE) 버튼을 눌렀을 경우 다이얼로그와 같은 추가 확인 절차 없이 즉각적으로 전체 문제 풀이가 종료되는 인터랙션 흐름 사용성 문제이다. 일반적으로 사용자는 시험 종료 조작 시 진행 상황 저장 여부나 종료 확인과 같은 추가 단계를 기대하지만, 해당 인터랙션에서는 이러한 절차 없이 과업이 즉시 종료되어 사용자의 심성 모델과 실제 인터랙션 전개 간에 불일치가 발생한다. 동영상 정보를 입력받았을 때 AI는 이를 사용성 문제로 진단하였다.

이러한 결과는 화면 스크린숏 기반 평가에서는 AI가 인터랙션 흐름 사용성 문제를 거의 탐지하지 못하지만 동영상 기반 평가는 일부 흐름 문제를 추가로 식별할 수 있었음을 시사한다. 하지만 인간 전문가가 지적한 인터랙션 흐름 전반을 판단하기에는 여전히 한계가 있다.

V. 결 론

이 연구는 AI가 모바일 앱 사용성 평가에서 인간 UX 전문가의 판단을 어느 수준까지 재현하거나 보조할 수 있는지를 검증하기 위해 저품질 UX 요소 탐지 실험과 실제 상용 앱을 대상으로 한 비교 평가를 수행하였다. 연구 결과 AI 기반 사용성 평가는 입력 방식에 따라 탐지할 수 있는 문제 유형과 한계가 명확히 구분되는 양상을 보였다.

먼저 화면 스크린숏을 입력으로 한 평가에서는 시각적 대비, 버튼 스타일, 레이아웃 구성 등 명시적이고 규칙화된 UI 요소에 대한 위반 탐지에서 비교적 안정적인 성능을 보였다. 이는 휴리스틱이나 플랫폼 가이드라인처럼 형식화된 규범을 기반으로 한 정적 UI 점검 영역에서 AI의 활용 가능성을 시사한다. 하지만 사용자의 조작에 따른 상태 변화나 인터랙션 흐름, 플랫폼 규범의 세부 구조, 그리고 문화적 사용성 판단이 요구되는 영역에서는 일관된 한계가 확인되었다. 명확한 규범 위반 요소를 ‘MZ 세대 트렌드’나 ‘행동 유도 효과’로 긍정 해석하며 문제로 인식하지 않았으며 인간 전문가가 지적한 모든 인터랙션 흐름 관련 사용성 문제를 탐지하지 못했다. 결과적으로 화면 스크린숏 기반 AI 평가는 정적인 UI 구조에 대한 표면적 해석에는 강점을 보이지만, 사용자의 실제 사용 행태, 사용 맥락, 기대된 심성 모델, 과업 수행 과정 전반이 복합적으로 얽힌 모바일 UX 문제를 인간 전문가와 같은 수준으로 종합적으로 판단하는 데에는 근본적인 제약이 있음을 보여준다.

동영상 기반 평가에서 인터랙션 흐름을 AI가 판단하여 일부 사용성 문제를 추가로 탐지할 가능성을 충분히 보였다. 다만 동영상 입력을 지원하지 않는 AI 도구가 존재하고 전반적인 탐지 성능에서 획기적인 개선을 보이지 않았으며 여전히 제한적이었다는 점에서 두 입력 방식의 비교에서는 이러한 제약을 함께 고려해야 한다. 동영상 입력 시 AI의 전반적인 사용성 문제 탐지 성능은 여전히 제한적이었다. 세 가지 LLM 모델 모두 의도적으로 설계된 저품질의 UX 문제를 탐지하지 못했으며 사용자 조작에 따른 상태 변화나 화면 전개 과정에서의 표현 일관성에 대한 정보 역시 보편적인 사용성 원칙, 플랫폼의 네이티브 규범, 문화적 관습에 근거한 종합적 판단으로 충분히 연결되지 못했다. 결과적으로 동영상 기반 사용성 평가는 화면 스크린숏 입력에 비해 인터랙션 흐름의 일부를 추가로 탐지할 수 있다는 장점이 있지만, 이를 바탕으로 전문가 수준의 포괄적인 사용성 판단에 이르기에는 한계가 존재한다.

이 연구는 AI 기반 사용성 평가에서 활용 방식과 한계를 실증적으로 확인했다는 점에서 의의를 갖는다. 시각적 대비, 버튼 스타일, 레이아웃 구성 등 명확하게 드러나는 UI 규칙 위반이나 가이드라인 준수 여부를 점검하고자 할 때에는 화면 스크린숏 기반 평가가 효율적인 보조 수단이 될 수 있다. 그러나 화면 스크린숏은 사용자 조작에 따른 상태 변화나 인터랙션 흐름과 같은 요소를 충분히 반영하지 못하므로 이러한 사용성 문제를 평가하고자 할 때에는 동영상 기반 입력을 활용하는 것이 상대적으로 효과적이다. 따라서 AI 기반 사용성 평가는 평가 목적에 따라 전략적으로 활용될 필요가 있음을 시사한다.

종합하면 현재의 AI 기반 사용성 평가는 인간 UX 전문가를 완전히 대체하기보다는 특정 문제 유형에 대해 인간 평가

자를 보완하는 것과 같은 한정적인 보조적 도구로 활용되는 것이 적절하다. 화면 스크린숏 기반 평가는 정적 UI 구조와 명시적 규칙 위반 탐지에 강점을 보였으며 동영상 기반 평가는 사용자 조작에 따른 상태 변화와 인터랙션 흐름을 반영할 수 있었다. 향후 AI가 인간 전문가 수준의 판단에 근접하기 위해서는 다중 화면에 걸친 인터랙션 흐름, 심성 모델에 기반한 사용성 파악, 문화적 규범 등을 통합적으로 추론할 수 있는 능력이 강화될 필요가 있다.

Ⅵ. 연구의 한계와 향후 연구 방향

이 연구는 단일 온보딩 프로토타입과 하나의 상용 앱을 대상으로 실험을 수행하였다는 점에서 한계를 가진다. 이러한 제한된 실험 환경은 연구 결과의 일반화 가능성을 확보하는 데 제약으로 작용할 수 있다. 향후 연구에서는 최소 5개 이상의 다양한 유형의 프로토타입과 상용 앱을 대상으로 실험 규모를 확장함으로써 AI 기반 사용성 평가 결과의 일반화 가능성을 보다 강화할 필요가 있다.

또한 인간 전문가 평가를 수행하는 과정에서 평가자 간 신뢰도(inter-rater reliability)를 측정하지 못했다는 점에서 한계를 가진다. 사용성 평가는 주관적 특성이 강하므로 평가자 간 일치도 검증이 중요하다. 향후 연구에서는 더 많은 UX 전문가를 참여시켜 평가 표본을 확대하고, Cohen's Kappa와 같은 평가자 간 신뢰도 지표를 활용하여 평가 기준의 일관성을 체계적으로 검증할 필요가 있다. 이를 통해 연구 결과의 신뢰성과 방법론적 엄밀성을 더 강화할 수 있을 것이다.

AI 모델 간 성능 차이에 대해 통계적 유의성 검증을 수행하지 못했다는 점에서 한계를 가지며 이에 따라 연구의 내적 타당성이 제한될 수 있다. 향후 연구에서는 입력 조건별 효과 등에 대한 통계적 유의성 검증을 통해 결과의 신뢰성을 확보할 필요가 있다.

또한 결론에서 제시한 이미지 대비 동영상 정보에서의 평가 역량 우수성은 두 조건에서 동일한 실험 조건으로 진행하는 데에 있어 일반화하기에는 제한적이었다는 점에서, 향후 연구에서는 동영상 입력을 지원하는 AI 도구를 대상으로 통제된 실험 조건을 구성하고 AI의 동영상 처리 메커니즘을 명확히 파악하여 입력 방식에 따른 효과를 체계적으로 검증할 필요가 있다.

감사의 글

이 연구 과제는 2025년도 교육부 및 강원특별자치도의 재원으로 강원 RISE 센터의 지원을 받아 수행된 지역 혁신 중심 대학 지원체계(RISE) 글로컬대학 30의 결과입니다(2025-RISE-10-009).

참고문헌

[1] S. Lubos, A. Felbering, G. Leitner, and J. Schwazer, "Towards Recommending Usability Improvements with Multimodal Large Language Models," arXiv:2508.16165, 2025. https://doi.org/10.48550/arXiv.2508.16165

[2] P. N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness,* Cambridge, MA: Harvard University Press, 1983.

[3] M. C. Trivedi and M. A. Khanum, "Role of Context in Usability Evaluations: A Review," arXiv:1204.2138, 2012. https://doi.org/10.48550/arXiv.1204.2138

[4] J. Nielsen. 10 Usability Heuristics for User Interface Design [Internet]. Available: https://www.nngroup.com/articles/ten-usability-heuristics/.

[5] Apple Inc. Human Interface Guidelines [Internet]. Available: https://developer.apple.com/design/human-interface-guidelines/.

[6] J. Liu, "AI in Automated and Remote UX Evaluation: A Systematic Review (2014-2024)," *Advances in Human-Computer Interaction*, Vol. 2025, 7442179, pp. 1-19, August 2025. https://doi.org/10.1155/ahci/7442179

[7] M. Kadegaonkar and K. Karim, "AI-Driven Usability Testing: Integrating Eye-Tracking Data and Agentic Systems for Automated UI Evaluation," in *Proceedings of the AAAI Summer Symposium Series (SuSS-25)*, Philadelphia: PA, pp. 244-254, 2025. https://doi.org/10.1609/aaaiss.v6i1.36059

[8] J. Díaz, C. Rusu, J. A. Pow-Sang, and S. Roncagliolo, "A Cultural-Oriented Usability Heuristics Proposal," in *Proceedings of the 2013 Chilean Conference on Human-Computer Interaction (ChileCHI '13)*, Temuco, Chile, pp. 82-87, 2013. https://doi.org/10.1145/2535597.2535615

[9] Y. W. Jeong and J. A. Kim, "Development and Cross-Cultural Validation of the Korean Version of Smartphone's Usability Heuristics (SMASH)," *Healthcare Informatics Research*, Vol. 23, No. 4, pp. 328-332, October 2017. https://doi.org/10.4258/hir.2017.23.4.328

[10] K. Richter, J. Nichols, K. Gajos, and A. Seffah, "The Many Faces of Consistency in Cross-Platform Design," in *Proceedings of the CHI 2006 Extended Abstracts on Human Factors in Computing Systems*, Montréal, Québec, Canada, pp. 1639-1642, April 2006. https://doi.org/10.1145/1125451.1125751

[11] S. H. Yoo and B. C. Hwang, "Mobile Users' Preference on UI Type of the Convergence Information Product - Based on the User Research Conducted in 4 Countries," *Archives*

of Design Research, Vol. 21, No. 1, pp. 73-82, February 2008.

[12] UX Pilot. UX Pilot - Superfast UX/UI Design with AI [Internet]. Available: https://uxpilot.ai/.

[13] D. A. Norman, *The Psychology of Everyday Things*, New York, NY: Basic Books, 1988.

[14] J. Nielsen and R. Budiu. Success Rate: The Simplest Usability Metric [Internet]. Available: https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/.

[15] B. Deka, Z. Huang, and R. S. Kumar, "ERICA: Interaction Mining Mobile Apps," in *Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology (UIST 2016)*, Tokyo, Japan, pp. 767-776, October 2016. https://doi.org/10.1145/2984511.2984581

[16] D. Saffer, *Designing for Interaction: Creating Innovative Applications and Devices*, 2nd ed. Berkeley, CA: New Riders, 2009.

원수정(Su Jeong Won)

2021년～현　재: 한림대학교 디지털인문예술전공
※관심분야 : UX 디자인

김성우(Sung Woo Kim)

2000년 : IDT(Info. Design & Tech.), Georgia Tech (석사)
2021년 : 연세대학교 정보대학원 디지털문화콘텐츠/UX 박사
2002년～2005년: 삼성전자 CTO 소프트웨어센터 인터랙션팀 선임
2007년～2009년: UX Manager, Philips Design Singpoare
2009년～2012년: KT 종합기술원 및 IPTV 사업본부 UX 매니저
2013년～2021년: 국민대학교 TED 경험디자인학과 부교수
2022년～현　재: 한림대학교 디지털인문예술전공 부교수
※관심분야 : UX, 경험 디자인, 사회혁신디자인