ISSN 2508-2817 KiDRS Korea Institute of Design Research Society 통권 39호, 2026. Vol.11, No.2

사용성 문제 발견 및 평가 신뢰성을 중심으로 본 휴리스틱 평가에서 인간 전문가와 대규모 언어모델(LLM)의 차이 분석

Understanding the Differences Between Human Experts and LLMs in Heuristic Evaluation : Focusing on Usability Issue Discovery and Evaluation Reliability

주 저 자 : 전서연 (Tian, Shu Yan) 국민대학교 TED 스마트경험디자인학과 석사과정

공 동 저 자 : 이정은 (Lee, Jung Eun) 국민대학교 TED 스마트경험디자인학과 석사과정

교 신 저 자 : 허정윤 (Heo, Jeong Yun) 국민대학교 TED 스마트경험디자인학과 교수
yuniheo@kookmin.ac.kr

https://doi.org/10.46248/kidrs.2026.2.598

접수일 2026. 06. 10. / 심사완료일 2026. 06. 13. / 게재확정일 2026. 06. 15. / 게재일 2026. 06. 30.

Abstract

This study compares the heuristic evaluation results of human experts and large language models (LLMs) to analyze common and unique findings by each entity and to examine the validity of the LLM evaluation results and the feasibility of proposed improvements. To this end, a prototype of a virtual train ticket reservation application was constructed incorporating a total of 40 Ground Truth Usability Issues based on Nielsen's 10 heuristics, and human experts and LLMs were instructed to independently perform heuristic evaluations according to the same criteria. The results showed that while both human experts and LLMs demonstrated detection rates of approximately 70% or higher, there were differences in the types of issues they discovered. While human experts primarily identified contextual problems requiring judgment based on actual usage experience and domain knowledge, LLMs were relatively better at detecting problems revealed through direct manipulation and checking for exceptions; integrating both results expanded the scope of detected issues to 92.5%. However, some problems reported by LLMs lacked sufficient grounds for judgment, and phenomena such as "hallucinations" and improvement proposals with low feasibility were also identified. These results suggest the possibility of a collaborative heuristic evaluation structure in which LLMs initially explore a broad range of problems, and human experts verify, interpret, and prioritize the results.

Keyword

Heuristic Evaluation(휴리스틱 평가), Large Language Model(대규모 언어모델), Human-AI Collaborative Evaluation(인간-AI 협업 평가)

요약

본 연구는 인간 전문가와 LLM의 휴리스틱 평가 결과를 비교하여 공통 발견 문제와 각 주체별 단독 발견 문제를 분석하고, LLM 평과 결과의 유효성과 개선안의 실행 가능성을 검토하는 것을 목적으로 한다. 이를 위해 Nielsen의 10가지 휴리스틱을 기준으로 총 40개의 의도적 사용성 문제를 삽입한 가상 열차 예매 애플리케이션 프로토타입을 제작하고, 인간 전문가와 LLM이 동일한 평가 기준에 따라 독립적으로 휴리스틱 평가를 수행하도록 하였다. 연구 결과, 인간 전문가와 LLM은 각각 대략 75% 이상의 탐지율을 보였지만 발견한 문제의 유형에는 차이가 있었다. 인간 전문가는 실제 사용 경험과 도메인 지식을 기반으로 판단해야 하는 맥락적 문제를 주로 발견한 반면, LLM은 직접 조작과 예외 상황 점검을 통해 드러나는 문제를 상대적으로 잘 발견하였으며, 두 결과를 통합 할 경우 발견 문제 범위는 92.5%까지 확장되었다. 다만 LLM이 보고한 일부 문제는 판단 근거가 충분하지 않았으며 환각 현상과 실행 가능성이 낮은 개선안도 확인 되었다. 이러한 결과는 LLM이 1차적으로 넓은 범위의 문제를 탐색하고, 인간 전문가가 그 결과를 검증·해석·우선순위화하는 협업적 휴리스틱 평가 구조의 가능성을 제안한다.

목차

1. 서론

1-1. 연구 배경

생성형 인공지능(Generative Artificial Intelligence, 이후 GenAI로 표기)과 대규모 언어모델(Large Language Model, 이후 LLM으로 표기)의 발전은 인간과 인공지능(AI) 간 협업 방식에 큰 변화를 가져오고 있다. 최근 LLM은 단순한 정보 생성 도구를 넘어 정보 탐색, 분석, 의사결정 지원, 창의적 문제 해결 등 다양한 인지 활동을 지원하는 협업 파트너로 활용되고 있다.

사용자 경험(User Experience, 이후 UX로 표기) 분야에서도 LLM의 활용 가능성이 확대되고 있다. 최근 연구들은 사용자 페르소나 생성,[1] 아이디어 발상 지원,[2] 사용자 리뷰 분석,[3] 디자인 피드백 생성,[4] 프로토타이핑 지원[5] 등 다양한 설계 활동에서 LLM의 활용 가능성을 보고하고 있다. Liu(2025)가 진행한 체계적 문헌고찰 연구에서도 머신러닝과 신경망을 넘어 LLM과 GenAI가 사용자 행동 모델링, 감성 분석, 자동 피드백 보고서 생성 등 다양한 UX 활동에 활용될 수 있음을 정리하였다.[6]

사용성 평가(Usability Evaluation)는 사용자 인터페이스의 문제를 발견하고 개선하기 위한 핵심 활동으로, 전통적으로 인간 전문가의 판단에 의존해 왔다. 이 중 휴리스틱 평가(Heuristic Evaluation)는 사전에 정의된 사용성 원칙에 따라 전문가가 인터페이스를 점검하여 잠재적인 사용성 문제를 식별하는 대표적인 사용성 평가 방법이다. 휴리스틱 평가는 소수의 평가자만으로도 비교적 적은 시간과 비용으로 설계 초기 단계의 문제를 효과적으로 발견할 수 있다는 장점이 있어 널리 활용되고 있다.[7] 이러한 장점에도 불구하고, 휴리스틱

1) Barambones, J., Moral, C., de Antonio, A., Imbert, R., Martínez-Normand, L., & Villalba-Mora, E., 'ChatGPT for learning HCI techniques: A case study on interviews for personas', IEEE, IEEE Transactions on Learning Technologies, Vol.17, 2024, pp.1460-1475.

2) Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T., 'Ideas are dimes a dozen: Large language models for idea generation in innovation', SSRN, Available at SSRN, No.4526071, 2023, pp.1-14.

3) Assi, M., Hassan, S., & Zou, Y., 'LLM-Cure: LLM-based competitor user review analysis for feature enhancement', ACM, ACM Transactions on Software Engineering and Methodology, Vol.35, No.4, 2026, pp.1-25.

4) Duan, P., Warner, J., Li, Y., & Hartmann, B., 'Generating automatic feedback on UI mockups with large language models', ACM, Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024. 05, pp.1-20.

5) Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M., 'Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale', arXiv, arXiv preprint arXiv:2306.01815, 2023.pp.1-5.

6) Liu, J., 'AI-Powered Automated and Remote UX Evaluation Methods: A Systematic Literature Review', ACM, Proceedings of the 43rd ACM International Conference on Design of Communication, 2025. 10, pp.10-16.

평가는 평가자의 전문성, 경험, 도메인 지식에 따라 결과가 달라질 수 있으며 분석 과정의 맥락 의존성으로 인해 결과의 재현 가능성에 제한이 있을 수 있다.[8)]

최근에는 LLM을 활용하여 사용자 인터페이스의 사용성 문제를 자동으로 탐지하고 개선안을 제안하는 연구들이 등장하고 있으며, 일부 연구에서는 인간 전문가와 유사한 수준의 문제 탐지 성능이 보고되고 있다.[9)] 이와 함께 LLM은 평가자의 주관성과 평가 과정의 높은 비용 및 시간 부담을 완화하고 보다 일관된 사용성 평가를 지원할 수 있는 자동화된 도구로 주목받고 있다.[10)]

이와 함께, 일부 연구에서는 인간 전문가가 놓친 문제를 인공지능이 발견하거나, 반대로 인공지능이 발견하지 못한 문제를 인간 전문가가 식별하는 사례도 보고되면서 두 평가 주체가 상호보완적 특성을 가질 가능성이 제기되고 있다.[11)]

그러나 기존 연구는 몇 가지 한계를 가진다.
첫째, 대부분의 연구가 발견된 문제의 수나 정확도와 같은 성능 지표에 집중되어 있어 인간 전문가와 LLM이 실제로 어떤 유형의 사용성 문제를 다르게 발견하는지에 대한 분석은 부족하다.[12)] 둘째, LLM이 발견한 사용성 문제의 신뢰성에 대한 검증이 충분히 이루어지지 않았다. 특히 문제의 유효성(Validity), 실행 가능성(Actionability) 측면에서 인간 전문가의 평가와 어느 정도 일치하는지에 대한 분석은 제한적이다. 셋째, 인간 전문가와 LLM의 평가 차이를 바탕으로 인간과 LLM간의 협업적 휴리스틱 평가를 어떻게 설계할 것인지에 대한 실증적 연구 역시 충분히 탐색 되지 않고 있다.[13)]

1-2. 연구 목적 및 연구 문제

본 연구는 휴리스틱 평가에서 인간 전문가와 LLM의 평가 결과를 비교 분석하고, LLM 기반 사용성 평가의 특성과 활용 가능성을 탐색하며, 이를 바탕으로 인간-인공지능 협업 휴리스틱 평가 (Human-AI Collaborative Heuristic Evaluation)의 가능성을 확인하는 것을 목적으로 한다.

구체적으로 첫째, 인간 전문가와 LLM이 발견한 사용성 문제를 비교하여 공통적으로 발견한 문제와 각각 고유하게 발견한 문제를 구분하고, 어떤 유형의 문제를 다르게 발견하는지 분석한다. 둘째, LLM이 발견한 사용성 문제의 유효성, 실행 가능성을 인간 전문가의 평가와 비교한다. 셋째, 인간 전문가와 LLM간 평가 차이를 바탕으로 협업적 휴리스틱 평가 설계를 위한 시사점을 도출하고자 한다. 이에 다음과 같이 연구 문제를 정의한다.

첫째, 인간 전문가와 LLM은 휴리스틱 평가에서 어떤 유형의 사용성 문제를 다르게 발견하는가?

둘째, LLM이 발견한 사용성 문제의 유효성, 실행 가능성은 인간 전문가의 평가와 차이가 있는가? 차이가 있다면 어떤 차이가 있는가?

셋째. 인간 전문가와 LLM의 평가 차이는 협업적 휴리스틱 평가 설계에 어떤 시사점을 제공하는가?

---

7) Fernández, J., & Macías, J. A., 'Heuristic-based usability evaluation support: A systematic literature review and comparative study', ACM, Proceedings of the XXI International Conference on Human Computer Interaction, 2021. 09, pp.1-9.

8) Hertzum, M., Molich, R., & Jacobsen, N. E., 'What you get is what you see: Revisiting the evaluator effect in usability tests', Taylor & Francis, Behaviour & Information Technology, Vol.33, No.2, 2014, pp.144-162.

9) Duan, et al., Op. cit. 2024, pp.1-20.

10) Hsueh, N. L., Lin, H. J., & Lai, L. C., 'Applying large language model to user experience testing', MDPI, Electronics, Vol.13, No.23, 2024, Article 4633, pp.1-19.

11) Guo, M., Li, J., Cheng, Y., Jin, Z., & Wang, L., 'AI UXpert: Comparing the performance and perception of AI and human experts in B2B ease of use evaluation', ACM, Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems, 2026. 04, pp.1-8.

12) Guerino, G., Rodrigues, L., Capeleti, B., Mello, R. F., Freire, A., & Zaina, L., 'Can GPT-4o evaluate usability like human experts? A comparative study on issue identification in heuristic evaluation', Springer Nature Switzerland, IFIP Conference on Human-Computer Interaction, 2025. 09, pp.381-402.

13) Rotaru, O., Orhei, C., & Vasiu, R., 'Hybrid usability evaluation of an automotive REM tool: Human and LLM-based heuristic assessment of IBM Doors Next', MDPI, Applied Sciences, Vol.16, No.2, 2026, Article 723, pp.1-40.

2. 관련 연구

2-1. 휴리스틱 평가와 평가자 효과

휴리스틱 평가는 Nielsen과 Molich(1990)가 제안한 대표적인 전문가 기반 사용성 평가 방법으로, 소수의 평가자가 사전에 정의된 사용성 원칙을 활용하여 인터페이스를 점검하고 잠재적인 사용성 문제를 발견하는 기법이다.[14] 사용자 테스트와 달리 실제 사용자를 모집하지 않고도 비교적 적은 비용과 시간으로 문제를 발견할 수 있다는 장점으로 인해 널리 활용되어 왔다. Nielsen의 10가지 사용성 휴리스틱은 현재까지 가장 보편적으로 사용되는 평가 기준으로 자리 잡고 있다.[15] 주로 웹 서비스, 모바일 애플리케이션, 스마트 디바이스 등 다양한 디지털 제품 및 서비스의 사용성 평가에 활용되고 있다.[16] 휴리스틱 평가는 설계 초기 단계에서 잠재적인 사용성 문제를 효과적으로 발견할 수 있는 실용적인 방법으로, 3~5명의 전문가만으로도 상당수의 사용성 문제를 발견할 수 있다.[17] 그러나 동일한 인터페이스를 평가하더라도 평가자마다 발견하는 문제가 서로 다른 평가자 효과(Evaluator Effect)가 반복적으로 보고되어 왔다. 이러한 차이는 평가자의 전문성, 경험, 도메인 지식의 차이에서 비롯되는 것으로 알려져 있다.[18] 발견된 문제의 심각도 평가(Severity Rating) 또한 평가자 간 일관성이 낮은 것으로 알려져 있다.[19] 이러한 결과는 휴리스틱 평가가 단순한 규칙 기반 점검이 아닌, 사용자의 목표와 과업 상황에 대한 평가자의 해석과 추론이 개입되는 과정임을 시사한다.

14) Nielsen, J., & Molich, R., 'Heuristic evaluation of user interfaces', ACM, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1990. 03, pp.249-256.

15) Nielsen, J., 'Enhancing the explanatory power of usability heuristics', ACM, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1994. 04, pp.152-158.

16) Quiñones, D., Rusu, C., & Rusu, V., 'A methodology to develop usability/user experience heuristics', Elsevier, Computer Standards & Interfaces, Vol.59, 2018, pp.109-129.

17) Nielsen, J., & Landauer, T. K., 'A mathematical model of the finding of usability problems', ACM, Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, 1993. 05, pp.206-213.

18) Hertzum, et al., Op. cit. 2014, pp.144-162.

19) Ibid., p.156.

2-2. 대규모 언어모델 기반 사용성 평가

최근 LLM의 발전은 사용성 평가 분야에도 새로운 가능성을 제시하고 있다. 최근 추론 성능이 향상된 ChatGPT, Gemini, Claude와 같은 멀티모달 대규모 언어모델은 사전에 정의된 휴리스틱을 기반으로 UI 화면 이미지를 분석하여 사용성 문제를 탐지하고 개선안을 제안할 수 있으며, 인간 전문가와 유사한 수준의 문제 탐지 성능을 보이는 것으로 보고되었다.[20]

Duan et al.(2024)은 GPT-4 기반 Figma 플러그인을 활용하여 사용자 인터페이스 목업에 대한 자동 휴리스틱 평가를 수행하였으며 텍스트의 명확성, 정보 구조의 일관성, 인터페이스 구성 오류 탐지 측면에서 유의미한 성과를 보고하였다.[21] Fernández와 Macías(2021)는 휴리스틱 평가 지원 연구에 대한 체계적 문헌고찰을 통해 자동화 도구와 평가 지원 시스템이 평가 효율성을 향상시킬 수 있음을 보고하였다.[22] 이러한 연구들은 LLM이 사용성 평가 과정에서 유용한 보조 도구로 활용될 가능성을 보여준다.

이후 연구들은 LLM을 활용한 휴리스틱 평가의 적용 범위를 확장하여 다중 화면 분석, 디자인 리뷰, UX 피드백 생성, 사용성 테스트 분석 등 다양한 평가 맥락에서 활용 가능성을 탐색하고 있다. 일부 연구는 LLM을 단순한 평가 도구가 아니라 UX 전문가의 평가 과정을 보조하는 지능형 평가 지원 도구로 활용할 수 있음을 제안하였다.[23]

그러나 현재 연구들은 주로 LLM과 인간 전문가 간의 문제 탐지 성능에 초점을 두고 있다.[24] 이들은 주로 발견된 문제 수, 정밀도(Precision), 재현율(Recall) 등의 정량적 지표를 중심으로 두 평가 주체를 비교해 왔다.[25] 그러나 인간 전문가와 LLM이 발견하는 문제의 유형적 차이에 대한 분석은 상대적으로 부족한 실

20) Zhong, R., McDonald, D. W., & Hsieh, G., 'Synthetic Heuristic Evaluation: A comparison between AI- and human-powered usability evaluation', arXiv, arXiv preprint arXiv:2507.02306, 2025, pp.1-21.

21) Duan, et al., Op. cit. 2024, pp.1-20.

22) Fernández, Macías, Op. cit. 2021, pp.1-9.

23) Kuang, E., 'Crafting Human-AI Collaborative Analysis for Usability Evaluations', Rochester Institute of Technology, 2025, pp.1-244.

24) Guerino, et al., Op. cit. 2025, pp.381-402.

25) Duan, et al., Op. cit. 2024, pp.1-20.

정이다.

LLM의 활용이 확대되면서 인공지능 기반 평가 결과의 신뢰성과 타당성에 대한 논의도 증가하고 있다. 사용성 평가에서 LLM의 제안 내용이 실제 개선안으로 반영되기 위해서는 해당 판단이 어떠한 근거에 기반하여 도출되었는지 함께 제공할 필요가 있다. 설명 가능한 인공지능(Explainable AI, 이후 XAI로 표기) 연구는 인공지능의 판단 근거를 사용자에게 제공함으로써 적절한 신뢰 형성을 지원할 수 있다고 제안한다.[26] 사용성 평가 맥락에서는 LLM이 실제 존재하지 않는 문제를 생성하는 환각(Hallucination) 현상,[27] 문제의 중요도를 과대 또는 과소 평가하는 현상,[28] 그리고 실행 가능성이 낮은 개선안을 제안하는 문제가 보고되고 있다.[29] 이러한 현상은 LLM이 발견한 사용성 문제가 실제로 유효한 문제인지, 사용자 경험에 얼마나 영향을 미치는지, 그리고 실제 설계 개선에 활용 가능한지에 대한 검증 필요성을 제기한다.

2-3. 인간 전문가와 LLM의 사용성 문제 발견 차이

휴리스틱 평가 연구에서 평가자 효과는 오랫동안 중요한 연구 주제로 다루어져 왔다. 인간 전문가는 개인의 경험과 도메인 지식에 따라 서로 다른 사용성 문제를 발견한다. 특히 숙련된 전문가일수록 사용자의 목표와 과업 흐름을 고려한 맥락적 문제를 발견하는 경향이 있다.[30] 반면 LLM은 휴리스틱 규칙을 일관되게 적용 할 수 있다. 하지만 현재의 평가 방식은 주로 정적 사용자 인터페이스 표현에 기반하기 때문에 사용자 행동 맥락이나 인터랙션 흐름을 충분히 고려하지 못할 가능성이 있다.[31]

그럼에도 불구하고 인간 전문가와 LLM이 어떤 유형의 사용성 문제를 다르게 발견하는지에 대한 체계적인 비교 연구는 상대적으로 부족하다. 기존 연구들은 주로 문제 발견 수나 정확도와 같은 정량적 성능에 집중되어 있으며, 발견된 문제의 유형 수준에서 차이는 아직 충분히 규명되지 못하였다.

2-4. 인간-인공지능 협업 기반 사용성 평가

최근 인간-인공지능 협업 연구는 인공지능을 인간 전문가의 대체재가 아니라 협업 파트너로 활용하는 방향으로 발전하고 있다.[32] UX 평가 분야에서도 인공지능이 잠재적인 사용성 문제를 탐지한 후, 인간 전문가가 이를 검토·보완하는 협업적 평가 방식이 제안되고 있다. Fan et al.(2022)은 사용성 테스트 분석 과정에서 인공지능이 설명 가능한 피드백을 제공할 경우 평가자의 문제 발견 성과와 신뢰가 향상될 수 있음을 보고하였으며 설명 제공이 인간과 인공지능 간 협업 품질 향상에 중요한 역할을 수행함을 제시하였다.[33]

그러나 현재까지의 연구는 주로 협업 과정 자체의 가능성을 탐색하는 수준에 머물러 있으며 인간 전문가와 인공지능이 실제로 어떤 유형의 문제를 다르게 발견하는지, 그리고 어떤 영역에서 서로를 보완할 수 있는지에 대한 실증적 분석은 부족하다.

3. 연구 방법

3-1. 연구 설계

본 연구는 인간 전문가와 LLM 기반 휴리스틱 평가 결과를 비교하기 위한 비교 실험(Comparative Evaluation Study)으로 설계되었다. 모든 평가 주체는 동일한 인터페이스와 평가 기준을 사용하여 독립적으로 평가를 수행하였다. 평가 대상은 의도적으로 사용성 문제를 삽입한 가상 열차 예매 애플리케이션 프로토타입이며 평가 집단은 UX 실무 경험 3년 이상의 인간 전문가 5인과 LLM 기반 휴리스틱 평가로 구성하였다. 인간 전문가 수는 Nielsen과 Molich(1990)의 연구를 참고해 일반적으로 권장되는 5인을 모집하였다.[34] LLM은 동일한 입력 조건에서도 서로 다른 결과를 생

26) Fan, M., Yang, X., Yu, T., Liao, Q. V., Zhao, J., 'Human-AI collaboration for UX evaluation: Effects of explanation and synchronization', ACM, Proceedings of the ACM on Human-Computer Interaction, Vol.6, No.CSCW1, 2022, pp.1-32.

27) Guerino, et al., Op. cit. 2025, p.397.

28) Zhong, et al., Op. cit. 2025, p.9.

29) Duan, et al., Op. cit. 2024, p.9.

30) Ko, J., Choi, J., Morales, C., Kim, D., & Ko, M., 'Criticmate: Stagewise Human-AI co-critique in single-screen UI evaluation', ACM, Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 2026. 04, pp.1-22.

31) Guerino, et al., Op. cit. 2025, p.397.

32) Guo, et al., Op. cit. 2026, p.1.

33) Fan, et al., Op. cit. 2022, pp.1-32.

34) Nielsen, Molich, Op. cit. 1990, p.255.

성하는 확률적 생성(Stochastic Generation) 특성을 고려하여 동일한 조건에서 5회 반복 평가를 수행하였다. 이를 통해 인간 전문가 집단과 동일한 기준에서 평가 결과를 비교하고, LLM 기반 휴리스틱 평가 결과의 일관성과 재현성을 검토하고자 하였다.

비교의 타당성을 확보하기 위해 평가 대상, 평가 기준, 평가 절차 및 산출물 형식을 통제하였다. 모든 평가 주체는 동일한 프로토타입과 동일한 휴리스틱 평가 체크리스트를 사용하였으며 문제 설명, 심각도, 개선 제안이 포함된 보고서 형식으로 결과를 기록하였다. 인간 전문가는 평가 결과를 공유하지 않은 상태에서 독립적으로 평가를 수행하였고 LLM 역시 각 회차를 독립 세션에서 수행하여 평가 결과 간 상호 영향을 배제하였다.

3-2. 평가 대상

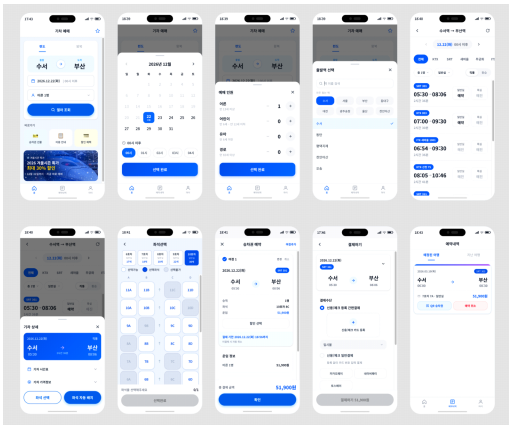
3-2-1. 평가 대상 선정

본 연구는 평가 주체 간 비교의 타당성을 확보하기 위해 의도적 사용성 문제(Ground Truth Usability Issues)가 삽입된 가상 열차 예매 애플리케이션 프로토타입을 평가 대상으로 선정하였다. 실제 운영 중인 서비스는 지속적인 업데이트와 개선 과정을 통해 주요 사용성 문제가 이미 해결되었을 수 있으며 서비스 업데이트로 인한 화면 변경, 개인화 기능, 시스템 오류 등 통제하기 어려운 변인이 존재한다. 이에 본 연구는 동일 조건에서 인간 전문가와 LLM의 평가 결과를 비교할 수 있도록 가상 프로토타입을 제작하였다.

평가 대상은 실제 사용 맥락을 반영하고자 국내 대표 열차 예약 서비스인 코레일톡과 SRT, 그리고 버스 및 SRT 예약이 가능한 티머니GO를 참고하여 프로토타입의 주요 화면 구조와 사용자 흐름을 설계하였다. 이를 통해 실제 사용자가 경험하는 서비스 환경과 유사한 맥락을 반영하고, 사회적으로 형성된 예약 서비스 이용 관행을 유지하고자 하였다.

프로토타입은 열차 예매 과정의 핵심 과업을 수행할 수 있도록 구성하였으며 홈, 조회 결과, 좌석 선택, 승차권 예약, 결제, 예약 내역의 6개 주요 화면과 역 선택, 날짜·시간 선택, 예매 인원 선택, 열차 상세정보 제공을 위한 4개의 하단 선택 패널(Bottom Sheet)을 포함한 총 10개 화면이 포함 되었다.

개발은 Figma를 활용하여 화면을 설계한 후 바이브 코딩(Vibe Coding) 방식을 활용하여 화면 간 전환과 주요 인터랙션이 동작하는 평가용 프로토타입으로 완성하였다. 최종 프로토타입의 주요 평가 화면은 [그림 1]에 제시하였다.



[그림 1] 가상 열차 예매 앱 프로토타입의 화면

또한 인간 전문가와 LLM의 문제 발견 능력을 비교하기 위해 프로토타입 내에 총 40개의 의도적 사용성 문제(Ground Truth Usability Issues)를 삽입하였다. 문제는 Nielsen의 10가지 휴리스틱을 기준으로 설계하였으며, 각 휴리스틱 유형이 고르게 포함될 수 있도록 구성하였다. 또한 선행 연구를 참고해 맥락적 고려가 필요한 사용성 문제와 화면상에서 바로 식별 가능한 사용성 문제가 균형 있게 분포되도록 하였다. 의도적 사용성 문제 설계에는 UX 연구자 3인(경력 20년, 5년, 3년)이 참여하였다. 초기 설계된 문제 목록에 대해 각 연구자가 독립적으로 검토를 수행하였으며, 세 연구자 모두가 사용성 문제로 판단한 항목만 최종 의도적 사용성 문제로 채택하였다. 또한 각 문제의 심각도는 세 연구자가 독립적으로 평가한 후 평균값을 산출하고 반올림하여 최종 심각도 점수로 사용하였다.

3-2-2. 인간 전문가 구성

인간 전문가 집단은 UX 실무 경력 3년 이상이며 휴리스틱 평가 경험을 보유한 5인으로 구성하였다. 이는 Nielsen과 Molich(1990)가 제안한 휴리스틱 평가의 적정 평가자 수를 고려한 것이다.[35] 평가자 직무는 다양하게 구성되어 나타났다. 평가에 참여한 전문가 구성은 [표 1]에 제시하였다.

35) Ibid., p.255.

[표 1] 인간 전문가 구성

ID	경력	직무
P1	5년차	프로덕트디자이너
P2	5년차	UX 디자이너
P3	4년차	UXUI 디자이너
P4	3년차	UX 디자이너
P5	3년차	서비스 UX 기획자

인간 전문가의 평가는 자료 전달, 개별 평가, 결과 회신의 순서로 진행하였다. 각 평가자는 동일한 프로토타입과 평가 체크리스트를 제공받은 후 독립적으로 평가를 수행하였고, 평가 과정에서 평가자 간 결과 공유나 연구자의 개입은 허용하지 않았다.

3-2-3. LLM 기반 휴리스틱 평가 구성

LLM 기반 휴리스틱 평가 (LLM Heuristic evaluation)는 주어진 목표에 따라 화면을 탐색하며 여러 단계의 작업을 순차적으로 수행하는 것을 필요로 한다. 이에 본 연구에서는 목표 지향적 과업 수행과 연속적인 작업 실행이 가능한 ChatGPT의 Agent Mode를 활용하여 구성하였다. LLM 기반 휴리스틱 평가의 구성 방식은 [그림 2]와 같이 4단계로 구성하였다. Agent Mode는 링크를 통해 여러 화면을 탐색할 수 있다. 이러한 특성은 기능 수행 과정에서 나타나는 화면 흐름 전반을 평가하는 데 적합하여 실제 사용자와 유사한 방식으로 평가가 가능하다.[36] 이에 본 연구에서는 프로토타입을 평가할 수 있도록 단계적 프로세스를 활용해 프롬프트를 구성하였다. LLM이 미구현 기능이나 시나리오 외 동작을 사용성 문제로 판단하지 않을 수 있도록 화면 탐색 가능 범위를 제공하여 해당 범위를 기준으로 사용성 평가를 진행할 수 있도록 하였다.

입력	1	**역할 부여**
		휴리스틱 평가 전문가 역할 부여
	2	**평가 기준 및 평가 방법 학습**
		• 10가지 휴리스틱 평가 원칙 • 심각도 평가 기준 (0-4점) 제공 • 화면 탐색 가능 범위 제공 • 프로토타입 URL • 평가 체크리스트
▼		
실행	3	**평가 진행**
		화면 진입 및 탐색 → 문제 식별 → 심각도 평가 → 개선 제안 작성 → 평가 내용 점검 → 결과 통합 표 작성
▼		
출력	4	**평가 결과 제공**

[그림 2] LLM 기반 휴리스틱 평가 수행 프로세스

3-3. 데이터 수집

본 연구는 인간 전문가 5인과 LLM 기반 휴리스틱 평가의 5회 반복 결과를 동일한 구조로 수집하였다.

평가를 위해 총 10개 화면에 걸쳐 86개의 평가 질문으로 구성된 휴리스틱 평가 체크리스트를 제공하였다. 체크리스트는 화면명, 화면 목적 및 역할, UI 영역, 휴리스틱 원칙, 평가 질문 등의 9개 항목으로 구성하였다.

평가자는 발견한 사용성 문제에 대한 문제 설명, 심각도, 개선 제안 등의 9개 항목으로 구성된 표에 기록하도록 하였다. 심각도는 Nielsen(1994)의 심각도 평가(Severity Rating)를 참고하여 0점(사용성 문제 없음)부터 4점(사용성 재앙 수준으로 즉각적인 개선이 필요한 문제)까지의 5단계 척도로 평가하였다.[37] 정해진 평가 기준 외에 참여자가 평가 과정에서 발견한 추가적인 사용성 문제를 기술할 수 있도록 개방형 응답 항목을 포함하였다.

수집된 데이터는 발견 문제(Issue), 문제 설명(Explanation), 개선 제안(Recommendation)의 세 가지 분석 단위로 구조화하였다.

발견 문제는 사전에 설계한 40개의 의도적 사용성

36) Lu, Y., Yao, B., Gu, H., Huang, J., Wang, Z. J., Li, Y., et al., 'UXAgent: An LLM agent-based usability testing framework for web design', ACM, Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025. 04, pp.1-12.

37) Nielsen Norman Group, How to Rate the Severity of Usability Problems, (2026.06.07.) www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems

문제를 기준으로 탐지 여부를 판정하였다. 또한 의도적 사용성 문제에 포함되지 않은 추가 발견 문제도 함께 포함하여 분석하였다.

4. 연구 결과

4-1. 인간 전문가와 LLM의 사용성 문제 발견 차이

4-1-1. 전체 탐지 결과

사전에 설계한 40개의 의도적 사용성 문제를 기준으로 인간 전문가와 LLM 기반 휴리스틱 평가의 탐지 여부를 분석하였다. 각 평가자가 기록한 문제 설명이 의도적 문제와 실질적으로 일치하는 경우 해당 문제를 탐지한 것으로 판단하였다. 이를 바탕으로 평가 주체별 탐지 여부를 O/X로 코딩하여 평가 주체별 탐지 결과를 정리하였으며, 분석 결과는 [표 2]에 제시하였다.

[표 2] 평가 주체별 의도적 사용성 문제 탐지 결과

	문제 정보			인간 전문가					LLM					총 발견 수
	발견 유형	원칙	문제요약	인간1	인간2	인간3	인간4	인간5	1회	2회	3회	4회	5회	
1	공통 발견	1. 가시성	예약·매진 상태의 배경과 글자 색상이 동일하여 좌석 상태를 빠르게 판단하기 어렵다.	O	O	O	O	O	O	O	O	O	O	10
2	공통 발견	10. 지원성	결제 실패 메시지가 화면 상단에만 표시되어, 실패 원인이나 고객센터 연락 방법을 충분히 확인하기 어렵다.	X	O	O	O	O	O	O	O	O	O	9
3	공통 발견	9. 회복 가능성	결제 완료 후 예약내역 화면에 "예약 취소" 외에 "좌석 변경" 같은 부분 수정 옵션이 없다.	O	O	O	X	O	O	O	O	O	O	9
4	공통 발견	4. 일관성	헤더의 출발지·도착지 표기가 "수서역 → 부산"처럼 "역" 접미사 사용에 일관성이 없다.	O	O	X	X	O	O	O	O	O	O	8
5	공통 발견	10. 지원성	결제 실패 시 실패 원인과 해결 방법을 안내하는 도움말을 쉽게 찾을 수 없다.	O	O	X	X	O	O	O	O	O	O	8
6	공통 발견	10. 지원성	예약 카드에 예약번호와 탑승 승강장 정보가 표시되지 않는다.	X	O	O	X	O	O	O	O	O	O	8
7	공통 발견	3. 통제성	인원 수의 최소·최대 범위가 표시되지 않아, 사용자가 조작 가능한 범위를 예측하기 어렵다.	X	X	X	O	O	O	O	O	O	O	7
8	공통 발견	2. 친숙성	결제 실패 안내 문구 "결제 실패되셨습니다"는 존댓말 표현이 어색하다. (올바른 표현: "결제가 실패했습니다")	X	X	X	O	O	O	O	O	O	O	7
9	공통 발견	5. 안전성	약관동의가 기본 체크되어 있고 필수/선택 구분이 없다.	X	X	O	O	X	O	O	O	O	O	7
10	공통 발견	4. 일관성	다른 바텀시트에는 상단 닫기 버튼이 있지만 날짜 선택 바텀시트에는 없어 구조적 일관성이 떨어진다.	X	O	X	O	O	O	X	O	X	O	6
11	공통 발견	3. 통제성	뒤로 가기 기능을 닫기(X) 아이콘으로 표시하여, 사용자가 화면 종료나 입력 삭제로 오인할 수 있다.	O	O	X	X	X	O	O	O	O	X	6
12	공통 발견	9. 회복 가능성	예약 상세 정보에 인원 수와 좌석 번호가 누락되어 있다.	X	X	O	O	X	O	O	O	O	X	6
13	공통 발견	8. 심미성	바로가기 카드의 배경색이 일관된 규칙 없이 사용되어, 시각적 균형과 통일감이 떨어진다.	O	X	X	X	X	O	O	O	O	O	6
14	공통 발견	6. 직관성	00시~04시만 노출되어, 더 많은 시간 옵션이 존재한다는 사실을 인지하기 어렵다.	X	O	X	O	O	X	X	X	O	O	5
15	공통 발견	6. 직관성	결제수단 분류와 표현이 직관적이지 않아 결제 방식을 빠르게 이해하기 어렵다.	O	X	X	O	O	X	O	X	X	O	5
16	공통 발견	2. 친숙성	결제 기한이 절대 시각으로만 제시되고 잔여 시간(예: "15분 이내")이 함께 안내되지 않아, 남은 시간을 직관적으로 파악하기 어렵다.	X	O	O	X	X	X	X	X	O	O	4
17	공통 발견	9. 회복 가능성	"좌석 자동 배치"만 강조되어, 좌석을 직접 선택하려는 사용자의 선택권이 충분히 드러나지 않는다.	X	X	X	O	X	O	O	O	O	X	5
18	공통 발견	8. 심미성	예약 취소 버튼이 시각적으로 강조되어, 핵심 정보인 예약 내역보다 먼저 시선을 끈다.	O	O	X	O	X	X	X	O	X	X	4
19	공통 발견	7. 효율성	출발역과 도착역 사이에 스왑(전환) 기능이 없어, 반대 방향 검색을 빠르게 수행하기 어렵다.	O	O	X	X	X	O	X	O	X	X	4
20	공통 발견	2. 친숙성	좌석의 방향(정·역방향)을 나타내는 시각적 표현이 제공되지 않는다.	O	O	X	O	X	X	X	X	O	X	4
21	공통 발견	4. 일관성	"승차권 예약" 타이틀이 단독으로 왼쪽에 정렬되어, 다른 화면의 헤더 정렬 규칙과 일치하지 않는다.	X	O	X	X	X	X	X	X	O	X	2
22	공통 발견	1. 가시성	결제 기한 안내의 시각적 위계가 낮고 총 결제 금액·CTA와 떨어져 있어 주목도가 떨어진다.	X	X	X	O	X	X	X	X	O	X	2
23	인간만	6. 직관성	열차 종류가 모두 같은 색으로 표시되어, 사용자의 비교와 선택에 도움이 되지 않는다.	O	O	O	O	O	X	X	X	X	X	5
24	인간만	6. 직관성	캘린더에서 토요일·일요일·공휴일이 평일과 같은 색으로 표시되어 구분되지 않는다.	O	O	X	O	O	X	X	X	X	X	4
25	인간만	1. 가시성	콘센트 석, 화장실, 유아 동반석 등의 표시가 제공되지 않는다.	O	O	O	X	O	X	X	X	X	X	4
26	인간만	7. 효율성	주요역 목록에 'ㄱㄴㄷㄹ' 인덱스가 제공되지 않아 빠른 탐색이 어렵다.	X	O	X	O	X	X	X	X	X	X	2
27	인간만	7. 효율성	조회결과 카드에 열차별 가격이 표시되지 않아, 가격 확인을 위해 상세 화면을 일일이 들어가야 한다.	X	X	X	O	X	X	X	X	X	X	1
28	인간만	2. 친숙성	홈 상단 별 아이콘의 기능과 즐겨찾기 대상이 불명확하다.	X	X	X	O	X	X	X	X	X	X	1
29	인간만	9. 회복 가능성	결제 실패 알림이 오류 상태처럼 강조되지 않아 실패 상황 인지가 약하다.	O	X	X	X	X	X	X	X	X	X	1
30	인간만	8. 심미성	홈 여정 입력 정보의 나열·표현·채도 때문에 선택값을 한눈에 파악하기 어렵다.	X	O	X	X	X	X	X	X	X	X	1
31	LLM만	1. 가시성	가로 스크롤 영역에 추가 옵션이 있음을 알리는 단서가 부족해 숨겨진 열차 종류를 인지하기 어렵다.	X	X	X	X	X	O	O	O	X	O	4
32	LLM만	5. 안전성	승객 수보다 적은 좌석을 선택해도 "선택 완료" 버튼이 활성화되어 예약 오류가 발생할 수 있다.	X	X	X	X	X	X	O	O	X	X	2
33	LLM만	10. 지원성	선택 불가 좌석을 탭해도 차단 이유나 상태 피드백이 제공되지 않는다.	X	X	X	X	X	X	X	O	O	X	2
34	LLM만	3. 통제성	하단 CTA 문구가 모호해 다음 단계 또는 결제 실행 결과를 예측하기 어렵다.	X	X	X	X	X	X	O	X	O	X	2
35	LLM만	7. 효율성	조회결과 목록의 추가 결과 탐색 가능성이 불명확하거나 스크롤이 제한된다.	X	X	X	X	X	X	O	X	O	X	2
36	LLM만	5. 안전성	"예약 취소"와 같은 액션 버튼이 주 행동 버튼과 가까이 배치되어, 사용자가 실수로 클릭할 가능성이 높다.	X	X	X	X	X	O	X	X	X	X	1
37	LLM만	4. 일관성	어린이 연령 기준 문구에서 '이상' 표현이 빠져 문법적으로 어색하다.	X	X	X	X	X	X	O	X	X	X	1
38	미발견	5. 안전성	현재 시간이 13:00이지만 오늘 날짜에서 00시 이후 시간을 선택할 수 있어, 이미 지난 시간대를 기준으로 조회·예매할 수 있다.	X	X	X	X	X	X	X	X	X	X	0
39	미발견	8. 심미성	어른·어린이·유아 항목 간 상하 여백이 10/11/12px로 불균형하다.	X	X	X	X	X	X	X	X	X	X	0
40	미발견	3. 통제성	"예약 취소" 액션의 적용 범위(여정 전체/좌석·시간 일부)가 명확하지 않아, 사용자가 수정 범위를 통제하기 어렵다.	X	X	X	X	X	X	X	X	X	X	0

모든 평가 주체가 탐지한 문제를 분석한 결과 5인의 인간 전문가는 총 30개(75.0%), 5회 평가한 LLM은 총 29개(72.5%)의 문제를 발견하였다. 인간 전문가와 LLM의 전체 탐지율은 유사한 수준으로 나타났으며, 두 평가 주체 모두 의도적 사용성 문제의 약 70% 이상을 발견하였다. 이 발견률은 기존 선행 연구와 거의 일치한다.

세부적으로 살펴보면, 인간 전문가와 LLM이 모두 탐지한 문제는 22개였으며, 인간 전문가만 탐지한 문제는 8개, LLM만 탐지한 문제는 7개, 양측 모두 탐지하지 못한 문제는 3개로 나타났다. 두 평가 주체 간 탐지 여부 차이를 검증하기 위해 continuity-corrected McNemar 검정을 수행한 결과, 통계적으로 유의한 차이는 나타나지 않았다($\chi^2 = 0.00$, $p = 1.00$). 이는 인간 전문가와 LLM이 전체적인 사용성 문제 검출 능력 측면에서 유사한 수준의 성능을 보였음을 의미한다.

개별 평가 기준으로 보면 인간 전문가는 평균 14.8개(범위 9-19개), LLM이 평균 18.2개(범위 15-21개)의 문제를 발견하였다. 그러나 반복 평가 결과를 통합하여 비교한 최종 탐지 범위는 두 집단 간 유사한 수준으로 나타났다.

종합하면, 인간 전문가와 LLM이 전체적인 사용성 문제 검출률 측면에서는 유사한 수준의 탐지 성능을 보였음을 시사한다. 즉, 본 연구에서 관찰된 인간 전문가와 LLM 기반 휴리스틱 평가 간의 차이는 단순한 검출 수의 차이라기보다 발견한 문제의 유형, 해석 방식, 그리고 문제의 특성에서 비롯될 가능성이 있음을 보여준다.

4-1-2. 발견 유형 분포

40개의 의도적 문제를 발견 주체에 따라 분류한 결과는 [표 3]과 같다. 공통 발견 문제는 22개(55.0%), 인간 전문가만 발견한 문제는 8개(20.0%), LLM만 발견한 문제는 7개(17.5%), 양측 모두 발견하지 못한 문제는 3개(7.5%)로 나타났다.

특히 인간 단독 발견 문제와 LLM 단독 발견 문제가 각각 8개와 7개로 유사한 규모를 보였다는 점은 전체 탐지율이 유사하더라도 실제로 탐지한 문제의 유형은 상당히 다를 수 있음을 보여준다.

인간 전문가와 LLM 기반 휴리스틱 평가 결과를 통합한 문제 발견 비율을 분석한 결과, 각 평가 주체가 평가를 독립적으로 수행했을때 보다 사용성 문제 발견의 범위가 확대 되는 것으로 나타났다. 구체적으로 미발견 문제를 제외하고 총 40개의 의도적 문제 중 37개의 문제가 발견되었으며, 이는 전체 사용성 문제의 총 92.5%에 해당 된다.

[표 3] 발견 유형별 의도적 문제 발견 비율

발견 유형	문제 수	비율
공통 발견	22	55.0%
인간 전문가 단독	8	20.0%
LLM 단독	7	17.5%
미발견	3	7.5%
합계	40	100%

4-1-3. 인간 전문가 단독 발견 문제

각 문제를 휴리스틱 유형과 문제 특성에 따라 범주화하여 두 평가 주체가 상대적으로 강점을 보이는 문제 유형을 비교하였다. 또한 인간 전문가와 LLM이 추가로 발견한 문제를 분석하여 어떤 유형의 문제를 새롭게 탐지하는지도 검토하였다.

인간 전문가만 발견한 문제는 총 8개이며, [표 4]에 제시하였다. 해당 문제들은 디자인 관습, 현실 세계와의 일치, 의사결정 지원, 사용 편의성, 정보 위계와 관련된 사용성 문제로 나타났다. 예를 들어 주말•공휴일 색상 구분 부재, 콘센트•유아석 등 정보 제공 부족, 열차 가격 미표시 등은 실제 사용 경험과 도메인 지식을 기반으로 판단해야 하는 문제였다. 이러한 문제들은 화면 자체의 결함보다 사용자의 기대와 서비스 이용 맥락을 고려할 때 비로소 문제로 인식된다는 특징을 보였다.

[표 4] 인간 전문가만 발견한 사용성 문제 유형

번호	분류	위반 원칙	문제 요약	발견 인원
1	직관적 상태 구분	6.직관성	열차 종류가 모두 같은 색으로 표시되어 구분되지 않는다.	5
2	디자인 관습	6.직관성	캘린더에서 토요일·일요일·공휴일이 평일과 같은 색으로 표시되어 구분되지 않는다.	4
3	현실 미반영	1.가시성	콘센트석·화장실·유아동반석 등의 정보가 제공되지 않는다.	4
4	편의성 고려	7.효율성	주요역 목록에 'ㄱㄴㄷㄹ' 인덱스가 제공되지 않아 빠른 탐색	4

			이 어렵다.	
5	의사결정 지원	7.효율성	조회결과 카드에 열차별 가격이 표시되지 않아 상세 화면을 일일이 확인해야 한다.	2
6	직관적인 기능 판별	2.친숙성	홈 상단 별 아이콘의 기능과 즐겨찾기 대상이 불명확하다.	1
7	직관적인 정보 파악	9.회복가능성	결제 실패 알림이 오류 상태처럼 강조되지 않아 실패 상황 인지가 약하다.	1
8	정보의 중요도별 분류, 시각적 위계	8.심미성	홈 여정 입력 정보의 나열·표현·채도 때문에 선택값을 한눈에 파악하기 어렵다.	1

4-1-4. LLM 단독 발견 문제

LLM만 발견한 문제는 총 7개이며, [표 5]에 제시하였다. 해당 문제들은 오류 예방, 상태 피드백, 시스템 반응, 시각적 단서 부족, 문장 표현 오류와 관련된 사용성 문제로 나타났다. 예를 들어 승객 수보다 적은 좌석을 선택해도 예약이 가능한 문제, 취소 버튼 배치로 인한 오클릭 가능성 등은 비정상적인 조작을 시도하는 과정에서 발견되는 문제였다. 이와 함께 가로 스크롤 영역의 탐색, 선택 불가 좌석의 반응 확인 등과 같이 직접 조작이 필요한 문제에서도 상대적으로 LLM이 많이 발견하였다. 이러한 문제들은 사용자의 기대와 서비스 이용 맥락 보다 화면 자체의 결함이나 조작 과정을 고려할 때 비로소 문제로 인식된다는 특징을 보였다.

[표 5] LLM 기반 휴리스틱 평가 과정에서 발견된 사용성 문제 유형

번호	분류	위반 원칙	문제 요약	발견 회차
1	사용자의 행동을 유도하는 시각적인 단서 부족	1.가시성	가로 스크롤 영역에 추가 옵션이 있음을 알리는 단서가 부족해 숨겨진 열차 종류를 인지하기 어렵다.	4회
2	오류 예방	5.안전성	승객 수보다 적은 좌석을 선택해도 "선택 완료" 버튼이 활성화되어 예약 오류가 발생할 수 있다.	2회
3	시스템 피드백 및 상태 인지	10.지원성	선택 불가 좌석을 탭해도 차단 이유나 상태 피드백이 제공되지 않는다.	2회
4	직관적인 기능 판별	3.통제성	하단 CTA 문구가 모호해 다음 단계 또는 결제 실행 결과를 예측하기 어렵다.	2회
5	편의성 고려	7.효율성	조회결과 목록의 추가 결과 탐색 가능성이 불명확하거나 스크롤이 제한된다.	2회
6	오류 예방	5.안전성	"예약 취소"와 같은 액션 버튼이 주 행동 버튼과 가까이 배치되어, 사용자가 실수로 클릭할 가능성이 높다.	1회
7	띄어쓰기, 맞춤법, 문장표현	4.일관성	어린이 연령 기준 문구에서 '이상' 표현이 빠져 문법적으로 어색하다.	1회

4-2. LLM 기반 휴리스틱 평가 결과 분석

4-2-1. 개선 제안의 실행 가능성

LLM이 발견한 29개의 문제를 대상으로 개선 제안의 실행 가능성을 평가하였다. 실행 가능성은 두 평가 주체가 제시한 개선안의 구체성과 실현 가능성을 기준으로 분석하였다. 이를 위해 UX 연구자 3인이 독립적으로 평가한 후 합의를 통해 최종 결과를 도출 하였다.

분석 결과 13개(44.8%)의 문제에서 LLM이 제안한 개선안은 실제 설계 개선에 반영하기 어려운 것으로 평가되었다. 특히 공통 발견 문제 11개에서 LLM이 문제 자체는 정확하게 식별하였으나, 제안된 개선안은 문제의 본질과 충분히 연결되지 않은 것으로 나타났다. 반면 동일 문제에 대해 인간 전문가가 제안한 개선안은 대부분 실제 설계 개선에 활용 가능한 수준으로 평가되었다. 실행 가능성의 분석 결과는 [표 6]에 제시하였다.

[표 6] LLM의 개선 제안 내용에 관한 실행 가능성 분석

발견 유형	발견 문제 수	AI 개선안	
		실행 가능	실행 어려움
인간+LLM	22개	11개	11개
LLM 단독	7개	5개	2개
전체	29개	16개	13개

4-2-2. 추가 발견 문제 분석

사전에 설계한 의도적 사용성 문제 이외의 평가 주체가 추가로 발견한 41개의 문제를 분석 하였다.
분석 결과 인간 전문가가 추가 발견한 문제 16개는 주로 화면 정렬, 아이콘 표현 등 구현 품질과 관련된 항목이 많았다. 그러나 발견된 문제들은 대부분 낮은 심각도 수준으로 평가되었다. 사용자 경험에 미치는 영향

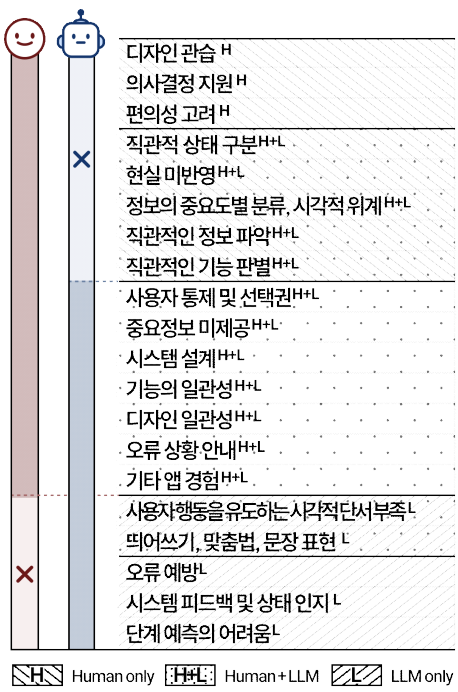
이 제한적인 것으로 해석할 수 있다. 반면 LLM이 추가 발견한 문제 21개는 실제 존재하지 않는 결함을 문제로 판단한 사례가 다수 포함되어 있었다. 특히 스와이프 조작이나 복합 인터랙션 관련 영역에서 이러한 현상이 반복적으로 나타났다. 이는 LLM의 평가 결과에 대해 추가적인 검증 절차가 필요함을 보여준다.

4-3. 인간 전문가와 LLM의 발견 특성 비교

결과를 종합하면 [그림 3]과 같이 인간 전문가와 LLM은 유사한 수준의 문제를 발견하였다. 하지만 각 주체가 발견한 문제의 유형과 평가 방식에는 차이가 존재하였다.

인간 전문가는 사용자 경험, 도메인 맥락, 사회문화적 관습에 기반한 문제를 상대적으로 잘 발견하였다. 반면 LLM은 오류 예방, 상태 전이, 시스템 반응과 같이 직접 조작과 예외 상황 점검이 필요한 문제를 상대적으로 잘 발견하였다. 명시적으로 확인 가능한 사용성 문제는 두 평가 주체 모두 높은 수준으로 발견하였다.

이러한 결과는 인간 전문가와 LLM이 서로 다른 강점과 한계를 가지며 상호보완적 관계를 형성할 수 있음을 보여준다.



[그림 3] 평가 주체별 사용성 문제 발견 특성 비교

5. 논의

5-1. LLM과 인간간의 차이와 특성

본 연구는 LLM 기반 휴리스틱 평가가 의도적 사용성 문제의 상당수를 탐지할 수 있음을 확인하였다. LLM은 40개의 의도적 문제 중 29개(72.5%)를 발견하였으며, 인간 전문가 집단의 발견율(75.0%)과 유사한 탐지율을 보였다. 특히 오류 예방, 시스템 피드백, 상태 인지, 시각적 탐색 단서와 같이 인터페이스를 직접 조작하거나 예외 상황을 시도해야 확인할 수 있는 문제를 효과적으로 발견하였다. 이러한 결과는 최근 LLM 기반 사용성 평가 연구가 제시한 자동 휴리스틱 평가의 가능성을 뒷받침한다.[38]

기존 연구가 주로 정적 UI 목업을 기반으로 문제 탐지 가능성을 검증하였다면, 본 연구는 Agent 기반 상호작용 환경에서 실제 조작을 수행하며 사용성 문제를 발견할 수 있음을 확인하였다는 점에서 의미가 있다.

그러나 본 연구는 LLM 기반 휴리스틱 평가의 한계도 확인하였다.

첫째, LLM은 실제 존재하지 않는 문제를 보고하는 환각을 보였다. 특히 스와이프, 바텀시트, 상태 전환과 같은 동적 인터랙션에서 이러한 현상이 반복적으로 나타났다. 이는 LLM이 인터페이스를 해석하는 과정에서 실제 시스템 상태와 내부 추론 결과를 혼동할 가능성이 있음을 시사한다.

둘째, LLM은 동일한 문제를 발견하더라도 인간 전문가보다 심각도를 낮게 평가하는 경향을 보였다. 특히 사용자 통제권, 정보 위계, 현실 세계와의 일치와 같은 문제에서 이러한 차이가 두드러졌다. 이는 LLM이 문제 존재 여부는 비교적 정확하게 식별하더라도, 해당 문제가 실제 사용자 경험에 미치는 영향을 판단하는 데에는 한계를 가질 수 있음을 의미한다.

셋째, LLM은 문제 발견에 비해 개선안 생성에서 상대적으로 낮은 성과를 보였다. 발견한 29개 문제 가운데 13개(44.8%)의 개선 제안은 실제 설계 개선안으로 활용하기 어려운 것으로 평가되었다. 이러한 결과는 LLM이 사용성 문제를 탐지하는 데에는 효과적일 수 있으나, 이를 실질적인 설계 개선안으로 발전시키는 데에는 한계가 있을 수 있음을 시사한다.

5-2. 평가자 효과의 확장: 인간 전문가와 LLM의 상이한 전략

38) Duan, et al., Op. cit. 2024, pp.1-20.

휴리스틱 평가 연구에서는 동일한 인터페이스를 평가하더라도 평가자마다 서로 다른 문제를 발견하는 평가자 효과가 반복적으로 보고되어 왔다.[39]

본 연구는 이러한 평가자 효과가 인간 평가자들 사이에서만 발생하는 현상이 아니라, 인간 전문가와 LLM이라는 서로 다른 평가 주체 간에도 나타날 수 있음을 보여준다.

인간 전문가는 디자인 관습, 의사결정 지원, 현실 세계와의 일치, 편의성 고려와 같이 사용자의 기대와 경험을 기반으로 판단해야 하는 문제를 상대적으로 많이 발견하였다. 반면 LLM은 오류 예방, 상태 피드백, 시스템 반응, 시각적 탐색 단서 부족과 같이 인터페이스를 직접 조작하거나 예외 상황을 검증하는 과정에서 드러나는 문제를 더 많이 발견하였다. 이러한 결과는 인간 전문가와 LLM이 서로 다른 문제를 발견하는 특성을 가지고 있음을 보여준다.

인간 전문가는 사용자의 경험과 맥락을 기반으로 문제를 해석하는 경향을 보였으나 LLM은 인터페이스의 구조와 휴리스틱 원칙 위반 여부를 체계적으로 점검하는 과정에서 문제를 발견하는 경향을 보였다. 이러한 차이는 두 평가 주체가 서로 다른 강점과 한계를 가진다는 것을 보여주고 사용성 평가에서 서로 상호보완적인 역할을 수행할 가능성을 시사한다. 또한 평가자 효과에 대한 논의를 인간 전문가 간 차이에서 인간-인공지능 간 차이로 확장할 수 있는 근거를 제공한다.

5-3. 인간-LLM 협업 평가 가능성과 설명 가능한 사용성 평가(XAI)의 필요성

본 연구 결과는 인간 전문가와 LLM이 대체 관계가 아닌 상호보완적 관계에 있음을 보여준다. 인간 전문가는 맥락적 해석과 우선순위 판단에 강점을 보였으며, LLM은 반복적인 탐색과 오류 기반 검증에 강점을 보였다. 또한 인간 전문가만 발견한 문제와 LLM만 발견한 문제가 각각 8개와 7개로 나타났다는 점은 두 평가 주체가 서로 다른 사각지대를 가지고 있음을 의미한다.

이러한 결과는 인간-인공지능 협업 연구가 제안하는 보완적 협업(Complementary Collaboration) 관점과도 연결된다. 보완적 협업 관점에서 인간-인공지능 협업은 두 주체의 서로 다른 강점을 결합 했을 때, 협업의 결과를 개선할 수 있다고 보았다.[40] 즉, 인간과 인공지능이 보유한 서로 다른 강점이 상호 보완적으로 결합될 때, 개별 주체가 단독으로 수행하는 경우보다 더 나은 결과를 도출할 수 있을 것으로 볼 수 있다. 사용성 평가의 목적이 단순히 많은 문제를 발견하는 것이 아니라 실제 개선 과제를 도출하는 데 있다는 점을 고려하면, 인간 전문가와 LLM을 결합한 협업 평가가 효과적인 대안이 될 수 있다.

이와 함께 본 연구는 사용성 평가에서 XAI의 중요성을 보여준다. LLM이 발견한 일부 문제는 판단 근거가 충분히 설명되지 않았으며, 실제로 존재하지 않는 문제를 보고하는 경우도 확인되었다. 이러한 결과는 AI가 제시한 평가 결과를 인간 전문가가 신뢰하기 위해서는 문제 자체뿐 아니라 문제를 판단한 근거 역시 함께 제공되어야 함을 의미한다.

선행 연구에서는 설명이 인간의 이해와 신뢰 형성의 핵심 요소임을 강조하였으며, 최근 인간-AI 협업 연구 역시 설명 가능성을 협업의 필수 조건으로 제시하고 있다.[41] 이때 형성되는 신뢰는 검토 가능하고 이의 제기를 할 수 있는 추론을 제시하는 것을 의미한다.[42] 이러한 맥락에서 신뢰는 AI의 결과를 무비판적으로 수용하는 것이 아니라 인간이 추론 과정과 판단 근거를 지속적으로 검토하고 평가할 수 있는 상태에서 형성된다.

향후 LLM 기반 사용성 평가는 단순히 문제를 발견하는 수준을 넘어, 휴리스틱의 어떤 항목을 위반 하였는지, 문제 판단의 근거는 무엇인지, 사용자에게 미치는 영향이나 해당 개선안을 제안한 이유 등을 함께 제시하는 설명 중심 평가 방식으로 발전할 필요가 있다. 이러한 접근은 사용성 평가에서 AI를 인간 전문가의 대체자가 아닌 신뢰 가능한 협업 파트너로 활용하기 위한 중요한 요건이 될 것으로 기대한다.

6. 결론 및 향후 연구

본 연구는 인간 전문가와 LLM이 수행한 휴리스틱 평가 결과를 비교하여 사용성 문제 발견 양상을 분석하였다.

분석 결과, 인간 전문가와 LLM는 전체 탐지율에서는 유사한 수준을 보였으나 발견한 문제의 유형에는

39) Hertzum, et al., Op. cit. 2014, p.154.

40) Kuang, Op. cit. 2025, pp.1-244.

41) Fan, et al., Op. cit. 2022, pp.1-32.

42) Ko, et al., Op. cit. 2026, pp.1-22.

차이가 있었다. 인간 전문가는 디자인 관습, 현실 세계와의 일치, 의사결정 지원과 같은 맥락적 문제를 상대적으로 잘 발견하였으며, LLM은 오류 예방, 상태 피드백, 시각적 탐색 단서와 같은 조작 기반 문제를 상대적으로 잘 발견하였다.

LLM이 사용성 문제를 발견하는 능력은 비교적 우수하였지만, 심각도 판단과 개선안 생성에서는 인간 전문가보다 낮은 수준의 신뢰성을 보였다. 환각 현상과 실행 가능성이 낮은 개선안을 보였는데 이러한 결과는 LLM 기반 평가가 여전히 인간 전문가의 검증을 필요로 함을 보여준다.

본 연구의 학술적 의의는 기존 연구가 주로 문제 발견 성능에 집중하였던 것과 달리, 인간 전문가와 LLM이 어떠한 유형의 사용성 문제를 다르게 발견하는지와 평가의 신뢰성을 함께 분석하였다는 점에 있다. 이와 함께 평가자 효과를 인간-LLM 관계로 확장하고, 인간-LLM 기반 인공지능 협업 평가 가능성을 실증적으로 제시하였다는 점에서도 의의를 가진다. 실무적 관점에서는 LLM이 먼저 1차 탐색을 수행하고 이후에 인간 전문가가 결과를 검증 하는 형태로 협업 평가 구조의 가능성을 제안하였다는 점에서 의미가 있다.

그러나 본 연구는 단일 도메인의 열차 예매 애플리케이션 프로토타입을 대상으로 수행되어 단일 LLM 모델과 특정 프롬프트 전략을 사용하였다는 한계를 가진다. 평가 기준 역시 특정 휴리스틱 프레임워크를 기반으로 구성되었으므로 결과를 일반화하는 데 주의가 필요하다.

향후 연구에서는 다양한 서비스 도메인과 UI 유형을 대상으로 분석을 확장하고 ChatGPT, Gemini, Claude 등 다양한 LLM 간 비교를 수행할 필요가 있다. 또한 인간 전문가와 AI Agent의 상호작용 과정을 포함하는 인간-LLM 기반 인공지능 협업 평가를 설계하고, 실제 평가 성과 향상 여부를 검증하는 연구가 요구된다. 더 나아가 설명 가능한 사용성 평가(XAI for UX Evaluation)를 구현하기 위한 프롬프트 구조와 인터페이스 설계 연구 역시 중요한 후속 과제로 제안한다.

참고문헌

1. Assi, M., Hassan, S., Zou, Y., 'LLM-Cure: LLM-Based Competitor User Review Analysis for Feature Enhancement', ACM Transactions on Software Engineering and Methodology, 2026.
2. Barambones, J., Moral, C., de Antonio, A., Imbert, R., Martínez-Normand, L., Villalba-Mora, E., 'ChatGPT for learning HCI techniques: A case study on interviews for personas', IEEE Transactions on Learning Technologies, 2024.
3. Duan, P., Warner, J., Li, Y., Hartmann, B., 'Generating automatic feedback on UI mockups with large language models', Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024.
4. Fan, M., Yang, X., Yu, T., Liao, Q. V., Zhao, J., 'Human-AI collaboration for UX evaluation: Effects of explanation and synchronization', Proceedings of the ACM on Human-Computer Interaction, 2022.
5. Fernández, J., Macías, J. A., 'Heuristic-based usability evaluation support: A systematic literature review and comparative study', Proceedings of the XXI International Conference on Human Computer Interaction, 2021.
6. Girotra, K., Meincke, L., Terwiesch, C., Ulrich, K. T., 'Ideas are dimes a dozen: Large language models for idea generation in innovation', SSRN, 2023.
7. Guerino, G., Rodrigues, L., Capeleti, B., Mello, R. F., Freire, A., Zaina, L., 'Can GPT-4o Evaluate Usability Like Human Experts? A Comparative Study on Issue Identification in Heuristic Evaluation', IFIP Conference on Human-Computer Interaction, 2025.
8. Guo, M., Li, J., Cheng, Y., Jin, Z., Wang,

L., 'AI UXpert: Comparing the Performance and Perception of AI and Human Experts in B2B Ease of Use Evaluation', Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems, 2026.

9. Hertzum, M., Molich, R., Jacobsen, N. E., 'What you get is what you see: Revisiting the evaluator effect in usability tests', Behaviour & Information Technology, 2014.
10. Hsueh, N. L., Lin, H. J., Lai, L. C., 'Applying large language model to user experience testing', Electronics, 2024.
11. Ko, J., Choi, J., Morales, C., Kim, D., Ko, M., 'Criticmate: Stagewise Human-AI Co-Critique in Single-Screen UI Evaluation', Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 2026.
12. Leiker, D., Finnigan, S., Gyllen, A. R., Cukurova, M., 'Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale', arXiv preprint, 2023.
13. Liu, J., 'AI-Powered Automated and Remote UX Evaluation Methods: A Systematic Literature Review', Proceedings of the 43rd ACM International Conference on Design of Communication, 2025.
14. Lu, Y., Yao, B., Gu, H., Huang, J., Wang, Z. J., Li, Y., Wang, D., 'UXAgent: An LLM agent-based usability testing framework for web design', Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems, 2025.
15. Nielsen, J., 'Enhancing the explanatory power of usability heuristics', Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1994.
16. Nielsen, J., Landauer, T. K., 'A mathematical model of the finding of usability problems', Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, 1993.
17. Nielsen, J., Molich, R., 'Heuristic evaluation of user interfaces', Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1990.
18. Quiñones, D., Rusu, C., Rusu, V., 'A methodology to develop usability/user experience heuristics', Computer Standards & Interfaces, 2018.
19. Rotaru, O., Orhei, C., Vasiu, R., 'Hybrid Usability Evaluation of an Automotive REM Tool: Human and LLM-Based Heuristic Assessment of IBM Doors Next', Applied Sciences, 2026.
20. Zhong, R., McDonald, D. W., Hsieh, G., 'Synthetic Heuristic Evaluation: A Comparison between AI- and Human-Powered Usability Evaluation', arXiv preprint, 2025.
21. Kuang, E., 'Crafting Human-AI Collaborative Analysis for Usability Evaluations', Rochester Institute of Technology, 2025.
22. www.nngroup.com