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

UX 휴리스틱 평가 도구로서의 LLM 활용 가능성과 한계

모델 간 심각도 평가 경향 차이를 중심으로

The Potential and Limitations of Using LLMs as UX Heuristic Evaluation Tools

Focusing on Differences in Severity Assessment Tendencies Across Models

주 저 자 : 정수빈 (Jeong, Su Been) 국민대학교 조형대학 AI디자인학과 학사과정

교 신 저 자 : 허정현 (Heo, Jung Hyun) 국민대학교 조형대학 AI디자인학과 교수
jh.heo@kookmin.ac.kr

https://doi.org/10.46248/kidrs.2026.2.132

접수일 2026. 05. 20. / 심사완료일 2026. 05. 29. / 게재확정일 2026. 06. 08. / 게재일 2026. 06. 30.

Abstract

Generative AI, particularly large language models (LLMs), is being actively applied across all stages of the design process, including UX usability evaluation. However, when using LLMs as evaluation tools, systematic differences in evaluation tendencies may arise depending on the model's specific training data and inference methods. This study empirically analysed the differences in severity judgements exhibited by three models—ChatGPT, Claude and Gemini—when evaluating the same 50 UI samples against Nielsen's 10 heuristics. Analysis using a mixed-effects model revealed that model type had a statistically significant effect on severity scores ($p < .001$), with a tendency to rate severity higher in the order ChatGPT (2.631) > Claude (2.417) > Gemini (2.031). The effect sizes between models ranged from moderate to very large ($d = 0.655$-1.837), and the coefficient of variation (CV < 1%) between repeated evaluations demonstrated that the judgement tendencies of each model were highly stable. This suggests that usability evaluations relying solely on a single LLM may be influenced by model-specific severity judgement tendencies. Therefore, it is recommended that cross-validation using multiple models or a parallel evaluation strategy involving human experts be applied when conducting LLM-based UX evaluations.

Keyword

Large language model(대규모 언어 모델), Heuristic evaluation(휴리스틱 평가), Usability evaluation (사용성 평가), Mixed-effects model(혼합 효과 모형)

요약

생성형 AI, 특히 LLM은 UX 사용성 평가 등 디자인 프로세스의 전반적인 단계에서 활발히 활용되고 있다. 그러나 LLM을 평가 도구로 사용할 경우, 모델 고유의 학습 데이터와 추론 방식에 따라 모델 간 평가 경향의 차이가 나타날 수 있다. 본 연구는 ChatGPT, Claude, Gemini 세 모델이 Nielsen의 10가지 휴리스틱을 기준으로 동일한 UI 50개를 평가할 때 나타나는 심각도 판단 차이를 실증적으로 분석하였다. 혼합 효과 모형 분석 결과, 모델 유형은 심각도 점수에 통계적으로 유의한 영향을 미쳤으며($p < .001$), ChatGPT(2.631) > Claude(2.417) > Gemini(2.031) 순으로 심각도를 높게 평가하는 경향이 확인되었다. 모델 간 효과 크기는 중간~매우 큰 수준($d = 0.655$~1.837)이었고, 반복 평가 간 변동계수(CV < 1%)는 각 모델의 판단 경향이 매우 안정적임을 보여주었다. 이는 단일 LLM에만 의존한 사용성 평가가 특정 모델의 심각도 판단 경향에 의해 영향을 받을 수 있음을 시사한다. 따라서 LLM 기반 UX 평가 시 복수 모델 교차 검증 또는 인간 전문가와의 병행 평가 전략을 적용할 것을 권장한다.

목차

1. 서론

1-1. LLM의 UX 활용 확산

최근 생성형 AI의 발전으로 인해 많은 디자인 분야에서 AI를 적극적으로 활용하고 있다. 이미지 생성, 3D 모델링, 영상 생성 등 다양한 AI가 디자인 전반에서 활용되고 있으나, 특히 LLM(large language model)은 디자이너의 브레인스토밍, 리서치, 평가 단계에서 널리 활용되고 있다. 2022년 공개된 openAI 사의 ChatGPT를 필두로 Google의 Gemini, Anthropic의 Claude 등 다양한 LLM이 디자인 과정 전반에서 사용되고 있으며, 높은 활용 가능성을 보인다. 특히 비용과 효율성의 측면에서 AI를 적극적으로 도입하고 있으며, UX 연구에도 이를 활용하는 움직임이 보인다. 예를 들어 금융 서비스 'Toss'는 디자이너들이 효율적으로 제품의 사용성을 점검할 수 있게 사용자처럼 학습된 AI '휴리봇'을 만들어 UX 리서치를 위해 자체적으로 사용하고 있다.[1] 이처럼 LLM 기반 사용자 테스트와 사용성 평가는 실무에서 비용 절감 효과를 기대할 수 있으나, 그 결과는 모델의 훈련 데이터와 알고리즘 구조에 따라 편향(bias)을 내포할 가능성이 있다.

1-2. 기존 LLM 편향 연구의 한계

LLM의 기존 편향 연구는 주로 성별, 인종, 정치 성향 등 사회적 편향을 중심으로 논의됐다. 그러나 최근 LLM의 활용 범위가 콘텐츠 생성, 의사결정 보조, 평가 및 분석 영역으로 확장됨에 따라, 편향 문제는 사회적 판단에만 국한되지 않는다. 특히 UX 디자인 분야에서 LLM이 사용성 평가, 디자인 피드백 등 전문가의 역할로 활용되는 사례가 증가하고 있어, LLM의 편향성은 디자인 품질을 평가한 결과 자체를 왜곡할 가능성을 내포한다. 그럼에도 불구하고 기존 연구들은 LLM이 UX 사용성 평가에서 어떠한 평가 경향을 보이는지, 그리고 LLM 모델 간 평가 결과에 유의미한 차이가 존재하는지 충분히 검증하지 않았다.

1-3. 연구 배경 및 목적

본 연구는 LLM의 평가 경향 차이를 'UX 사용성 평가 맥락에서 관찰되는 모델 간 평가 경향 차이(inter-model evaluative tendency difference)'로 정의한다. 이는 인간 전문가 기준 대비 정확도 오류나 사회적 편향과는 구분되며, 동일한 UI에 대해 서로 다른 LLM이 휴리스틱 기준을 적용하는 방식에서 나타나는 체계적인 판단 차이를 의미한다. 특정 LLM이 일부 휴리스틱 항목에서 문제 판단 기준을 상대적으로 더 엄격하거나 느슨하게(strictness, leniency) 설정하는 경향으로 나타날 수 있다. 이러한 정의는 모델 간 평가 경향 차이를 단순한 환각(hallucination)이나 우연이 아닌, 모델 고유의 평가 경향(systematic tendency)으로 간주한다.

본 연구가 다루는 '모델 간 평가 경향 차이' 개념은 통계학적 점 추정 오차나 사회적 편향, 그리고 인간 전문가 기준선과의 정확도 오류와는 구분된다. 이는 동일한 평가 과업을 수행한 LLM 모델들 사이에서 체계적으로 관찰되는 판단 경향의 분산(inter-model variance in judgment tendency)을 의미한다. 즉, 본 연구가 검증하는 것은 "어떤 모델이 옳은가?"가 아닌 "모델 간에 체계적인 평가 경향 차이가 존재하는가?"이며 이는 평가자 간 신뢰도(inter-rater reliability) 연구에서 평가자 간 분산을 측정 대상으로 삼는 전통과 맥을 같이 한다. 선행 연구는 학습 데이터의 구성과 분포, 문화적 배경의 차이가 모델의 판단 기준 형

1) 토스테크, 휴리봇 이야기 #1: 토스는 AI 봇에게 사용자 인터뷰를 한다,(2026.05.12.) toss. tech/article/research-platform-AI

성에 유의미한 영향을 미친다고 보고해 왔다.[2][3] 이러한 선행 연구의 논의를 바탕으로, 본 연구는 학습 데이터 세트의 차이가 모델 간 평가 편향을 설명할 수 있는 하나의 가능 요인일 수 있다고 가정하되, 그 원인을 직접 규명하기보다 동일한 UI에 대한 휴리스틱 평가 결과가 LLM 모델별로 어떠한 차이를 보이는지를 비교 분석하는 데 초점을 둔다. 이를 통해 LLM을 평가 도구로 활용하는 UX 디자인 실무 및 연구 환경에서 모델 선택이 평가 결과에 미치는 영향을 실증적으로 제시하고, LLM 기반 사용성 평가의 신뢰성과 한계에 대한 논의를 확장하는 데 기여하고자 한다.

1-4. 연구 방법 및 범위

본 연구는 LLM을 UX 사용성 평가의 평가자로 간주하고, 동일한 UI에 대한 휴리스틱 평가 결과를 비교·분석하는 비교 관찰 연구(comparative observational study)로 설계되었다. 연구의 목적은 LLM 간 평가 결과의 차이와 반복적으로 관찰되는 경향을 분석하는 데 있으며, 특정 평가의 원인을 인과적으로 규명하는 데 있지는 않다. 연구 가설과 분석 지표 및 방법은 [표 1]과 같다. 연구는 공개된 UI 화면과 LLM 출력을 분석 대상으로 하며, 개인 식별 정보를 포함하지 않는다. 또한 본 연구는 평가 결과의 차이와 경향을 분석하는데 초점을 두며, 이러한 차이가 발생한 원인을 모델의 학습 데이터나 내부 구조 차원에서 규명하지는 않는다. 또한 평가 결과의 차이를 중심으로 분석하며, 문제 탐지 민감도나 평가 결과의 방향성에 대한 분석은 향후 연구 과제로 남긴다.

[표 1] 연구 가설과 분석 지표 및 방법

연구 가설	분석 대상 지표	지표 정의 및 산출 방식	분석 방법	해석 범위
H1. LLM 모델에 따라 휴리스틱 평가 결과에는 통계적으로 유의미한 차이가 존재할 것이다.	전체 평균 심각도 점수	UI별·모델별 심각도 점수의 평균값	혼합효과 모형	특정 LLM이 전반적으로 엄격하거나 관대한 평가를 수행하는지 확인
H1a. H1의 모델 효과는 개별 휴리스틱 항목 수준에서도 관찰될 것이다.	휴리스틱 항목별 심각도 점수	각 UI에 대해 LLM이 부여한 심각도 점수를 휴리스틱 항목별로 평균화	혼합효과 모형 (Fixed: LLM 모델 / Random: UI)	UI 단위 반복 측정 구조를 고려하여, UI 수준 변이를 반영한 상태에서 모델 간 평가 차이를 검증
H2. 동일 LLM 모델 내 반복 실행 간 심각도 판단은 높은 안정성을 유지할 것이다.	반복 실행(run) 간 변동성	5회 반복 실행에서 세션 평균 심각도의 표준편차 및 변동계수	기술통계 (Run SD, CV)	각 모델의 판단 경향이 실행 간 안정적으로 재현되는지 확인

2. 이론적 배경

선행 연구에 따르면, LLM은 훈련 데이터에 내재한 사회적 편향(social bias)을 비판 없이 재생산하거나 증폭시킬 수 있다는 우려가 광범위하게 제기되고 있다.[4] 이러한 연구 동향은 크게 세 가지 흐름으로 나타난다.

2-1. 편향의 정량적 탐지 및 교차 편향 분석 심화

초기 연구는 LLM이 인종, 종교, 성별, 연령 등 다양한 영역에서 편향을 내포하고 있음을 확인하고, 이를 수학적 개념 모델을 통해 시스템적으로 정량화하려는 프레임워크(framework)를 개발하는 데 집중했다. 선행 연구(방준성 외, 2023)는 LLM을 사용하는 AI 챗봇과 인간 간의 대화 인터랙션 과정에서 발생하는 편향성 검사 및 평가를 위한 프레임워크 개발 방법을 논의하고, 편향성 완화의 필요성을 제기하였다.[5] 이후 연구의 초점은 단순한 단일 도메인 편향을 넘어, 현실 세계의 복잡성을 반영하는 복합적인 교차 편향(intersectional bias) 분석으로 심화되었다.

선행 연구(이주은·배호, 2025)는 기존 연구의 제한적인 데이터 세트 규모와 단일 도메인 편향 측정의

2) 김유진, 강조은, 김한샘, '언어모델의 편향 개선을 위한 프롬프트 엔지니어링 연구: ChatGPT를 활용한 정치인 감성분석 말뭉치를 중심으로', 한국코퍼스언어학회, Corpus Linguistics Research, Vol.8, No.1, 2023.06, pp.49-66.

3) 이주은, 배호, '합성 데이터를 활용한 Large Language Model의 인종 및 성별 교차 편향 측정 벤치마크', 한국정보처리학회, 정보처리학회 논문지, Vol.14, No.2, 2025.02, pp.82-94.

4) 김현웅, 안현철, '대규모 언어 모델(LLM) 응답에 내재된 미묘한 성차별에 대한 연구', 한국빅데이터학회, 한국빅데이터학회지, Vol.10, No.1, 2025.06, pp.91-107.

5) 방준성, 이병탁, 박판근, '대규모 언어 모델을 사용하는 인공지능 기반 대화형 챗봇의 편향성 평가 프레임워크 개발 방법', 한국방송•미디어공학회, 방송공학회 논문지, Vol.28, No.5, 2023, pp.545-563.

한계를 극복하기 위해, 합성 데이터 생성(DPSDA2) 기법을 활용하여 총 16,082개의 대규모 벤치마크 데이터 세트를 구축함으로써 LLM의 인종 및 성별 교차 편향을 정량적으로 평가할 수 있는 확장 가능하고 포괄적인 프레임워크를 제시하였다. 이처럼 심화된 분석을 통해, LLM이 통계적 사실을 넘어 사회적 고정 관념적 응답(pro-stereotype)에 크게 치우치는 경향이 있음을 실증적으로 확인하였다.[6] 또한, 편향의 유형 역시 노골적인 차별을 넘어 '미묘한 성차별(gender microaggressions)'처럼 간접적이고 암묵적인 형태에 집중하여 분석되었다. 선행 연구(김현웅 외, 2025)는 한국(Naver CLOVA X), 미국(OpenAI ChatGPT-4.0), 중국(DeepSeek V3)의 대표적인 LLM들을 대상으로 응답에 내재된 미묘한 성차별(gender microaggressions) 표현을 실증적으로 비교 분석하여, 세 모델 모두 통계적으로 유의미한 수준의 성별 편향을 포함하고 있음을 입증하였다. 또한 해당 연구는 AI 평가자와 인간 평가자의 이중 평가 체계를 도입함으로써, 문화적 특성을 고려한 LLM 편향 분석을 위한 체계적인 프레임워크와 윤리적 정렬의 필요성을 제시하였다.[7]

2-2. 편향 완화를 위한 프롬프트 엔지니어링 활용

편향을 탐지하는 연구와 함께, LLM 출력의 편향을 개선하고 완화하기 위한 방법론 연구도 활발하게 이루어졌다. 특히, LLM의 출력 품질을 조절하는 핵심 기술인 프롬프트 엔지니어링(prompt engineering)이 편향 개선의 실질적인 방법론으로 주목받았다. 선행 연구(김유진 외, 2023)는 ChatGPT를 활용하여 국내 정치인에 대한 감성분석을 수행함으로써 언어모델이 진보 진영에 대해 긍정적으로 편향되어 있음을 정량적으로 확인하였다.[8] 선행 연구(김민채·박상희, 2025)는 LLM의 높은 활용 가능성에도 불구하고 디자인 프로세스에 관한 구체적인 연구가 부족하다는 문제의식에서 출발하여, 디자인씽킹 핵심 요소를 기준으로 프롬프트 엔지니어링 주요 기법을 체계적으로 분석하였다.[9] 연구들은 프롬프트 엔지니어링 기법을 체계적으로 적용함으로써, 언어모델의 편향된 출력물이 중립적인 결과물로 유의미하게 변화될 수 있음을 제시하였다. 이는 디자이너가 LLM을 활용하는 과정에서 프롬프트를 조작(예: 역할 부여, 답변 조건 명시) 함으로써 모델의 편향성을 의도적으로 제어할 수 있는 가능성을 시사한다.

2-3. LLM 기반 평가 시스템의 윤리적 한계

이처럼 LLM의 활용 가능성이 증대하고 평가 도구로 사용되려는 움직임(예: UX 리서치)이 나타나고 있음에도 불구하고, LLM 기반 평가 결과는 내재된 편향으로 인해 여전히 윤리적 한계를 가진다. LLM들은 고정관념을 반박하려는 노력(anti-stereotype)을 보일지라도, 특정 집단에 대한 선택이 과도하게 많아지는 역편향(overcorrection bias) 현상까지 보고되었다.[10] 결국, 사용자 테스트나 사용성 평가와 같은 UX 연구 분야에서 LLM 기반 서비스를 사용할 때, 사용자들은 답변의 유용성과 정확성만큼이나 '참여와 윤리'를 포함한 서비스의 안전성과 신뢰성 확보가 가장 중요한 요소라고 평가한다. 선행 연구(이서후·김승인, 2023)는 ChatGPT 3.5와 네이버 Clova X라는 대표적인 LLM 기반 AI 서비스들을 대상으로, '피드백과 표현', '직관적인 디자인', '맥락과 목적', '참여와 윤리', '자연스러움'의 다섯 가지 테마를 활용한 사용자 경험(UX) 평가 프레임워크를 제시하고 실증적으로 비교 분석하였다. 이러한 연구는 LLM이 훈련 데이터에 기반한 편향을 재생산하고 개인 정보 활용 등에 대한 명확한 안전장치를 제공하지 못할 경우, 이는 LLM 기반 평가 결과의 신뢰성을 저해하는 핵심 요인이 될 수 있음을 시사한다.[11]

2-4. UX 휴리스틱 평가의 특성과 평가자 간 신뢰도

선행 연구(Nielsen & Molich, 1990)에 따르면, 휴리스틱 평가(heuristic evaluation)는 전문가 기반 사용성 평가 방법으로, 평가자가 사전에 정의된 원칙(휴리스틱)을 기준으로 UI의 사용성 문제를 식별하고 심각도를 판단하는 방식이다. 이 방법은 사용자 참여

6) 이주은, 배호, pp.82-94.

7) 김현웅, 안현철, pp.91-107.

8) 김유진, 강조은, 김한샘, pp.49-66.

9) 김민채, 박상희, '디자인 프로세스 수행 효율성 향상을 위한 LLM 프롬프트 엔지니어링 기법 연구-디자인씽킹을 중심으로-', 한국브랜드디자인학회, 브랜드디자인학연구, Vol.23, No.1, 2025.03, pp.5-22.

10) 이주은, 배호, pp.82-94.

11) 이서후, 김승인, 'LLM기반AI 서비스사용자경험비교연구: ChatGPT와 Clova X를 중심으로', 한국상품문화디자인학회, 상품문화디자인학연구, No.75, 2023, pp.11-22.

없이 비교적 적은 비용으로 사용성 문제를 탐지할 수 있다는 장점으로 인해 실무와 연구 환경에서 널리 활용되고 있다. 그러나 선행 연구는 개별 평가자의 문제 탐지 결과가 본질적으로 불안정하며, 복수의 평가자 결과를 합산해야 적정 수준의 탐지율을 확보할 수 있음을 실증하였다. 이는 평가자 효과(evaluator effect)가 휴리스틱 평가 방법론에 본질적으로 내재해 있음을 의미하며, 평가자의 전문성, 경험, 그리고 휴리스틱 기준에 대한 해석 방식에 따라 문제 탐지 결과와 심각도 판단이 달라질 수 있다는 점은 오래전부터 지적됐다.[12]

이처럼 휴리스틱 평가가 평가자 의존적 속성을 지닌다는 사실은 평가자 간 신뢰도(inter-rater reliability) 문제로 직결된다. 평가자 간 신뢰도란 서로 다른 평가자가 동일한 대상을 평가할 때 얼마나 일관된 판단을 내리는지를 측정하는 개념으로, 평가자 기반 연구에서 결과의 타당성을 확보하기 위한 핵심 지표로 기능한다. LLM을 평가자로 활용할 때도 이러한 평가자 간 신뢰도 관점을 동일하게 적용할 수 있다. 즉, LLM 모델마다 다른 문제를 보고하고 다른 심각도를 부여하는 현상은 LLM 고유의 특수한 문제가 아니라, 휴리스틱 평가 방법론에 내재된 평가자 효과가 LLM 평가자에게도 동일하게 적용되는 것으로 해석할 수 있다. 본 연구는 이 관점에서 LLM 모델 간 판단 경향 차이를 평가자 간 분산(inter-rater variance)의 맥락으로 분석한다.

2-5. LLM 기반 사용성 평가 연구 동향

최근 생성형 AI의 발전과 함께 LLM을 UX 사용성 평가 도구로 활용하려는 시도들이 증가하고 있다. 선행 연구(김민채·박상희, 2025)는 디자인씽킹 핵심 요소를 기준으로 LLM 프롬프트 엔지니어링 기법을 체계적으로 분석하여 LLM이 디자인 프로세스 전반에서 활용 가능성을 가지면서도, 출력 품질이 프롬프트 설계 방식에 따라 크게 달라짐을 보고하였다.[13] 이러한 흐름 속에서 LLM을 평가자로 직접 투입하는 실증 연구도 등장하고 있다. 선행 연구(Zhong 외, 2025)는 멀티모달 LLM을 활용한 휴리스틱 평가를 수행한 후 그 결과를 숙련 UX 실무자의 평가와 비교하여, AI가 인간 평가자보다 더 높은 문제 탐지율을 보이면서도 특정 UI 컴포넌트 인식 실패와 화면 간 위반 탐지의 약점을 동시에 나타냄을 확인하였다.[14] 이는 AI 평가자의 성과가 문제의 총량이 아닌 탐지 특성의 차원에서 평가되어야 함을 시사한다. 이러한 선행 연구들은 LLM이 평가 도구로서 가능성을 갖지만, 모델 고유의 판단 경향이 평가 결과에 구조적 영향을 미칠 수 있음을 시사한다. 본 연구는 이러한 흐름 위에서 복수의 LLM 모델 간 판단 경향 차이를 UX 휴리스틱 평가 맥락에서 실증적으로 비교함으로써 기존 논의를 확장하고자 한다.

2-6. AI 평가자의 한계

LLM을 평가 도구로 활용할 때 핵심적으로 고려해야 할 문제는 판단의 일관성(consistency)과 모델 간 경향 차이(inter-model variance)이다. LLM의 출력은 동일한 입력에 대해서도 확률적 생성 과정으로 인해 미세한 변동을 보일 수 있으며, 이는 평가 안정성(evaluation stability)에 영향을 미친다. 선행 연구(Liu 외, 2023)는 LLM을 텍스트 생성 결과의 평가자로 활용하는 G-Eval 프레임워크를 제안하며, LLM 평가자가 자신과 유사한 산출물에 더 호의적인 경향을 보일 수 있다는 고유 선호 구조의 가능성을 명시적으로 지적하였다.[15] 이는 UX 평가 맥락에서도 모델별 고유의 내부 전략이 심각도 판단에 영향을 미칠 수 있음을 시사한다.

나아가 서로 다른 LLM 모델을 동일한 평가 과업에 투입했을 때 모델 간 결과 차이가 실질적으로 발생한다는 점은 최근 연구에서 직접적으로 확인되고 있다. 선행 연구(Campos 외, 2025)는 GPT-4o와 Gemini를 인간 전문가와 함께 사용성 검사에 투입하여 비교시 AI 평가자가 인간이 놓친 결함을 발견하는 보완적 가치를 가지면서도 거짓양성 증가와 중복 보

12) Nielsen, Jakob & Molich, Rolf, 'Heuristic Evaluation of User Interfaces', ACM, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems(CHI '90), 1990.04, pp.249-256

13) 김민채, 박상희, pp.5-22.

14) arXiv, Synthetic Heuristic Evaluation: A Comparison between AI- and Human-Powered Usability Evaluation, (2026.05.31.) arxiv.org/abs/2507.02306

15) Liu, Yang, Iter, Dan, Xu, Yichong, Wang, Shuohang, Xu, Ruochen & Zhu, Chenguang, 'G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment', Association for Computational Linguistics, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing(EMNLP 2023), 2023.12, pp.2511-2522

고라는 한계를 함께 나타냄을 확인하였으며, 인간과 AI를 결합하였을 때 최종 평가 성능이 가장 높았다고 보고하였다.[16] 이는 LLM을 독립적 평가자로 활용하기보다 인간 전문가를 보완하는 병렬 평가자로 설정하는 전략의 타당성을 지지한다. 또한 국내 연구(이주은·배호, 2025)는 합성 데이터를 활용한 대규모 벤치마크를 통해 서로 다른 LLM 모델이 동일한 평가 과업에서 상이한 편향이 보임을 실증적으로 확인하였다.[17] 이러한 연구들은 단일 LLM에 의존한 평가 결과가 해당 모델의 고유한 판단 경향을 반영할 수 있으며, 복수 모델 비교를 통한 교차 검증의 필요성을 공통으로 뒷받침한다.

3. 실험 설계 및 결과

본 절에서는 LLM이 Nielsen의 휴리스틱을 기반으로 UI를 평가할 때 나타나는 판단 경향성과 일관성을 분석하기 위한 실험 설계 전반을 기술한다. 단일 실행에 의존하는 기존 연구의 한계를 극복하고자, 동일 조건에서 복수의 반복 실행(repeated runs)을 수행하는 방식을 채택하였다. 이하에서는 실험 목적 및 설계, 반복 실행 구조, 프롬프트 설계, 통제 요인의 순으로 설계 세부 사항을 설명한다.

3-1. 실험 목적 및 설계

본 실험은 LLM이 Nielsen의 10가지 사용성 휴리스틱을 기반으로 UI를 평가할 때 나타나는 판단 경향성과 일관성을 분석하기 위해 반복 실행(repeated runs) 기반 실험 설계를 적용하였다. LLM은 동일 입력에 대해서도 확률적 생성 과정을 통해 미세한 판단 변동을 보일 수 있으며, 이러한 특성은 평가 안정성(stability)과 일관성(consistency) 측면에서 중요한 분석 대상이 된다. 이에 본 실험은 다음 두 가지 연구 목적을 중심으로 설계되었다.

1. 동일 UI에 대해 LLM은 평균적으로 어떠한 휴리스틱을 선택하며, 어느 수준의 심각도(severity)를 부여하는가? (판단 경향성 분석)

2. 동일 조건에서 반복 실행 시 LLM의 판단은 얼마나 일관되게 유지되는가? (판단 변동성 분석)

이를 위해 각 LLM은 동일한 50개의 UI 세트에 대해 총 5회 반복 실행되었다.

3-1-1. UI 샘플 구성

본 연구는 도메인별 특성 차이를 분석하는 것을 주요 목표로 하지 않고, UX 사용성 평가에서 빈번히 활용되는 표준 UI 패턴의 다양성을 확보하는 방향으로 실험 샘플을 구성하였다. 이에 따라 입력 중심 UI, 목록 탐색 UI, 상태 피드백 UI, 설정 화면, 정보 전달 중심 UI 등 다양한 인터페이스 유형을 포함하여 총 50개의 테스트 샘플을 구성하였다. UI 샘플 선정 기준 및 출처는 하단의 [표 2]와 같다.

[표 2] UI 샘플 선정 기준 및 출처

구분	내용
선정 목적	UX 사용성 평가에서 빈번히 활용되는 표준 UI 패턴의 다양성 확보
선정 기준	Nielsen Norman Group의 UI Elements Glossary를 바탕으로 상호작용 목적에 따라 5개 범주로 재범주화
샘플 출처	공개된 UI 화면(앱 스토어 스크린샷, UI 디자인 참조 사이트 등)
제외 기준	특정 도메인에 편중되지 않도록 단일 서비스·브랜드 UI의 과다 포함 배제
맥락 시나리오	각 UI의 주요 사용자 목적과 전형적 사용 상황을 연구자가 직접 작성하여 프롬프트에 포함
총 샘플 수	50개

UI 유형 분류는 Nielsen Norman Group의 UI Elements Glossary[18] 에서 제시하는 UI 요소들을 바탕으로, 연구자가 상호작용 목적에 따라 5개 범주로 재범주화하여 구성하였다. 각 범주는 다음과 같이 정의된다. Data input은 Text box, Listbox, Toggle 등 사용자가 시스템에 정보를 입력하거나 값을 선택하는 요소를, Navigation은 Navigation Bar, Tab Bar, Bottom Sheet 등 화면 간 이동 및 콘텐츠 탐색을 지원하는 요소를 포함한다. System feedback은 Dialog, Snack bar, Spinner, Progress Bar 등 시스템 상태와 처리 결과를 사용자에게 전달하는 요소를, Configuration은 Toggle, Segmented Button 등 시스

16) arXiv, Will AI also replace inspectors? Investigating the potential of generative AIs in usability inspection, (2026.05.31.) arxiv.org/abs/2510.17056

17) 이주은, 배호, pp.82-94.

18) Nielsen Norman Group, User-Interface Elements: Glossary,(2026.05.12.) www.nngroup.com/articles/ui-elements-glossary/

템 설정 및 옵션 조정에 관련된 요소를, Information display는 Card, Carousel, Calendar Picker 등 콘텐츠와 정보를 구조화하여 표시하는 요소를 포함한다. 이는 휴리스틱 평가가 사용자-시스템 상호작용 맥락을 분석 단위로 하기 때문이며, 각 UI는 주요 상호작용 목적이 단일하지 않을 수 있음을 반영하여 복수 라벨 방식으로 매핑하였다. 또한 각 UI는 모바일 및 데스크톱 환경을 포함하며, 모든 UI에는 1, 2문장의 사용 맥락 시나리오를 제공하여 평가자가 단순한 시각적 판단이 아닌 실제 사용 상황을 고려하여 휴리스틱 평가를 수행할 수 있도록 하였다. 사용 맥락 시나리오는 각 UI의 주요 사용자 목적과 전형적인 사용 상황을 간결하게 기술하는 방식으로 연구자가 직접 작성하였으며, UI별 기능적 맥락과 사용자 행동 흐름을 고려하여 개별적으로 구성하였다. 작성된 시나리오는 LLM에 전달되는 프롬프트 내에 UI 이미지와 함께 포함되어, 모델이 평가 수행 시 해당 맥락 정보를 참조할 수 있도록 설계하였다. [표 3]은 UI 샘플 유형 분류 및 구성을 나타내며, 각각의 UI가 복수의 카테고리에 중복으로 카운팅(counting) 되었다. 사용 맥락 시나리오의 작성 예는 하단의 [표 4]와 같다.

[표 3] UI 샘플 유형 분류 및 구성

Interaction Category	UI Samples	모바일	데스크탑
Data input	20개	18개	2개
Navigation	21개	20개	1개
System feedback	18개	17개	1개
Configuration	6개	6개	0개
Information display	19개	18개	1개

[표 4] 사용 맥락 시나리오의 작성 예시

UI_ID	UI 유형	플랫폼	사용 맥락 시나리오
GU01	로그인 입력 화면	모바일	사용자는 기존 계정으로 로그인하려 한다. 현재 화면에서 입력 필드 구성, 로그인 버튼의 배치 및 활성 조건, 그리고 계정 관련 보조 기능을 확인할 수 있다.
BU11	로딩 스피너만 존재하는 화면	모바일	사용자는 작업 요청 직후 대기 중이며, 시스템이 정상 처리 중인지 판단해야 한다.
GU13	파일 업로드 선택 모달	데스크톱	사용자는 업로드할 파일을 선택하려 한다. 드래그 앤드 드롭 또는 파일 선택 버튼 중 어떤 방식을 사용할지 판단한다.

3-1-2. 반복 실행 구조 및 run 정의

본 연구에서 run은 'UI 전체 세트를 한 번 평가한 실행 단위'로 정의하였다. 즉, 특정 모델에 대해 run=1은 50개의 UI sample 전체를 독립적으로 평가한 결과를 의미하며, run=2는 동일 조건에서 전체 UI sample set를 다시 평가한 결과를 의미한다. 각 모델은 run=1부터 run=5까지 총 5회 반복 실행되었다. 각 UI 평가는 독립 세션에서 수행되었다. 동일 세션 내에서 다수의 UI를 연속적으로 평가할 경우, 이전 UI에 관한 판단 맥락이 이후 평가에 영향을 미칠 수 있다. 이러한 판단 맥락의 오류로 인한 오염을 방지하기 위해 본 연구는 'UI 1개 = 1세션' 원칙을 적용하였다. 모든 run에서 UI 순서는 동일하게 유지되었으며, 실행 환경 역시 변경되지 않았다.

이와 같은 설계를 통해 run 간 차이는 프롬프트 변경이나 입력 조건의 차이가 아닌, 모델의 확률적 생성 특성에서 기인하는 것으로 해석될 수 있도록 하였다.

3-1-3. LLM 평가 절차

본 연구에서는 ChatGPT는 GPT-5.2 (OpenAI), Gemini는 3.1 Pro(Google), Claude는 Sonnet 4.6 (Anthropic)을 LLM 평가 대상으로 선정하였다. 모든 모델에는 동일한 평가 프롬프트가 적용되었으며, 프롬프트에는 Nielsen의 10가지 휴리스틱 기준과 심각도 척도를 명시하였다. 각 UI에 대해 LLM 모델별로 5회 반복 평가를 수행하여 평가 결과의 변동성과 일관성을 동시에 관찰하였다. 반복 평가를 통해 산출된 결과는 평균값을 중심으로 분석되었다. LLM 평가의 일관성을 확보하기 위해 구조화된 고정 프롬프트(structured fixed prompt template)를 설계하였다. 프롬프트는 다음과 같은 계층적 구조를 포함하며, 구조화된 고정 프롬프트 설계 요소는 [표 5]와 같다. 이러한 구조는 자연어 변동을 최소화하고, 자동 전처리 및 long-format 변환을 용이하게 하기 위한 통제 장치로 기능하였다.

[표 5] LLM 휴리스틱 평가를 위한 구조화된 고정 프롬프트 설계 요소

설계 요소	내용
1. 역할 고정	모델은 ' UX Expert conducting a formal heuristic evaluation study' 로 역할을 고정하였다. 이는 모델의 응답을 비공식적 대화가 아닌 통제된 연구 상황으로 정렬하기 위한 cognitive anchor로 작용한다.
2. 평가 과제 명시	모델은 UI Image, UI Type (메타데이터, 판단 기준에 영향 금지), Platform, Usage context scenario의 입력 요소를

	기반으로 평가하도록 지시되었다.
3. 휴리스틱 평가 기준 고정	Nielsen의 10가지 휴리스틱(H1-H10)을 명시적으로 제공하였다.
4. 심각도 척도 고정	Severity는 1-4 범위로 정의되었으며, 각 수준의 의미를 명확히 제시하였다.
5. 엄격한 지침	* 문제 1개 = 1 분석 단위 * 단일 휴리스틱만 선택 * 유사 문제 병합 * 최대 3개 문제까지 보고 * 긍정적 피드백 금지 * 구조화된 출력 형식 외 설명 금지
6. 출력 형식 강제화	Problem 1 / Heuristic ID / Severity / UI evidence 형식을 수정 불가 조건으로 고정하였다.

3-1-4. 통제 요인

반복 실행 간 비교가능성을 확보하기 위해 다음 요소를 통제하였다. 이를 통해 run 간 변동은 외생적 요인이 아닌 모델 내부 생성 특성에 기인하도록 설계하였다. 실험 통제 요소와 적용 내용은 [표 6]와 같다.

[표 6] 실험 통제 요소 및 적용

통제 요소	통제 내용
프롬프트 내용 고정	모든 모델 및 반복 실행에 동일한 프롬프트 내용을 적용하여 지시 방식의 차이가 결과에 영향을 미치지 않도록 하였다.
휴리스틱 정의 고정	Nielsen의 10가지 휴리스틱 원칙(H1-H10)을 프롬프트 내에 명시적으로 제공하여, 모델이 동일한 평가 기준을 참조하도록 하였다.
심각도 척도 고정	1-4점의 4점 척도 및 각 점수에 대한 정의를 프롬프트에 명시하여 모든 평가에 동일한 기준이 적용되도록 하였다. (0은 문제없음이므로, 문제가 있을 시 보고를 하도록 설계하였다.)
출력 형식 고정	평가 결과의 구조적 일관성을 확보하기 위해 고정된 텍스트 출력 형식(문제 번호 / 휴리스틱 ID / 문제 요약 / 심각도 / UI 근거)을 지정하였으며, 형식 외 설명 출력을 금지하였다.
UI 입력 자료 고정	모든 모델에 동일한 UI 이미지 및 사용 맥락 시나리오를 제공하여 평가 대상의 차이를 배제하였다.
UI 제시 순서 고정	순서 변동을 배제하기 위해 모든 반복 실행에서 UI 제시 순서를 동일하게 고정하였다.
모델 설정 및 실행 환경 고정	Temperature 등 생성 관련 파라미터와 API 호출 환경을 고정하여 실행 환경 차이로 인한 변동을 최소화하였다.

3-2. 평가 출력의 코딩 및 데이터 전처리

LLM의 평가 출력은 사전에 정의된 코딩 규칙에 따라 정량화되었다. 코딩 규칙은 다음과 같다. 첫째, 하나의 문제는 하나의 분석 단위로 간주한다. 둘째, 동일한 문제의 반복 서술은 의미적 유사성을 기준으로 병합한다. 셋째, 하나의 문제에 복수의 휴리스틱 항목이 적용될 수 있는 경우, 가장 핵심적인 휴리스틱 항목 하나만을 할당한다. 각 문제는 휴리스틱 항목 ID, 문제 요약, 심각도 점수(1-4), 문제 근거로 구성된 구조화된 형식으로 정리되었으며, 긍정적 평가 서술은 정량 분석에서 제외하였다. 모든 모델의 평가 결과에는 동일한 코딩 규칙이 적용되었다. LLM의 응답은 사용자와의 이전 대화 맥락에 영향을 받을 수 있다는 점을 고려하여, 본 연구에서는 각 UI에 대한 평가를 독립된 세션(session)에서 수행하였다. 즉, UI 1개당 별도의 채팅창을 생성하고, 해당 세션에서는 단일 UI에 대한 평가만을 수행하였다. 동일 UI에 대한 반복 평가(총 5회) 또한 매회 새로운 세션에서 독립적으로 실행하였다. 또한 이전 세션 기록 저장 혹은 맥락(context)을 반영하는 내부 설정이 있을시, 해당 기능을 off 상태로 전환하였다. 이를 통해 이전 평가 응답이 이후 평가 판단에 영향을 미치는 잠재적 맥락의 누적을 최소화하고, 각 평가를 통계적으로 독립된 관측치로 간주할 수 있는 실험 조건을 확보하였다. 코딩 작업은 제1 저자가 단독으로 수행하였으며, 사전에 정의된 코딩 규칙을 기반으로 파이썬(python) 스크립트를 활용하여 전처리 및 정량화 작업을 자동화하였다. 의미적 유사성에 의한 병합 판단은 동일한 UI 요소에 대해 동일한 문제 원인을 지칭하는 경우를 병합 대상으로 정의하였으며, 표현 방식의 차이가 있더라도 지칭 대상과 근거가 동일한 경우 하나의 분석 단위로 처리하였다.

LLM 평가 결과는 '문제 1개 = 1행'의 long-format으로 변환되었다. 각 행에 포함된 정보는 [표 7]과 같다. 문제없음으로 판단된 UI 역시 명시적으로 기록하여, 문제 탐지율 분석 시 표본 왜곡이 발생하지 않도록 하였다. 또한 UI 단위 분석을 위해 파생 변수를 생성하였다. UI 단위 파생 변수는 [표 8]과 같다. 또한 데이터의 신뢰성을 확보하기 위해 다음 절차를 적용하였다. 데이터 무결성 검증 및 오염 방지 절차는 [표 9]와 같다.

[표 7] Long-format 데이터 구조

분석 단위	각 행에 포함된 정보
문제 단위	model_id
	run_id
	ui_id
	problem_index(P1-P3)
	heuristic_id
	severity
	issue_description

[표 8] UI 단위 파생 변수

분석 단위	생성된 파생 변수
UI 단위	problem_count (UI 당 문제 개수)
	mean_severity
	severity_sd (run 간 표준편차)
	yes_ratio (5회 중 문제 판단 비율)
	heuristic_mode (최빈 휴리스틱)

[표 9] 데이터 무결성 검증 및 오염 방지 절차

통제 요소	통제 내용
1. 형식 검증	heuristic_id는 H1-H10 범위로 제한하였으며, severity는 1-4 범위를 벗어나지 않도록 자동 검증하였다.
2. 논리 일관성 검증	문제 여부와 problem 블록값 간의 논리적 일관성을 점검하였다.
3. 세션 독립성 확보	각 UI는 독립 세션에서 평가되었다.
4. 맥락 오염 방지 지시문 포함	" Evaluate independently. Do not assume any evaluation patterns from previous tasks. "라는 지시문을 포함하였다.
5. 비가시 정보 추론 금지	UI 이미지 및 시나리오에 명시되지 않은 기능을 추론하지 않도록 명시하였다.

3-2-1. 분석 전략

LLM 반복 평가 결과는 다음 세 관점에서 분석되었다. 첫째, 중심 경향 분석에서는 UI별 평균 심각도 및 문제 개수를 산출하였다. 선행 연구에서 반복 실행 결과의 최솟값, 중앙값, 최댓값 및 spread를 핵심 지표로 설정하여 중심 경향과 분포 특성을 함께 분석하는 것이 LLM 판단 경향을 체계적으로 이해하는 데 필수적임을 강조한 바 있다.[19] 둘째, 변동성 분석에서는 run 간 심각도 표준편차 및 휴리스틱 선택 일치율을 산출하였다. 복수의 run에 걸친 모델 출력의 일관성 측정은 LLM의 불확실성 추정을 위한 확립된 방법으로, 모델이 동일 입력에 대해 일관된 판단을 산출하는지를 평가하는 핵심적 분석 전략으로 선행 연구에서 보고되었다.[20] 셋째, 필요시 혼합 효과 모형을 적용하여 UI 및 run을 랜덤 효과로 설정하고 모델 효과를 추정하였다.[21][22] 이 구조는 단일 실행 결과에 의존하지 않고 LLM의 판단 경향성과 안정성을 동시에 분석할 수 있도록 설계되었다는 점에서, LLM을 평가자로 활용하는 연구에서 신뢰도와 타당도 분석을 위한 체계적인 심리측정학적 프레임워크 적용의 필요성을 강조한 선행 연구의 권고 내용과 부합한다.[23] run 간 변동성은 3.5절에서 별도로 분석하였으며, run 간 CV가 1% 미만으로 나타나 random 효과로의 추가 포함은 불필요한 것으로 판단하였다. 마지막으로 본 연구에서 휴리스틱 심각도 점수(1~4점)는 통계학적 분류상 순서형 자료(Ordinal Data)에 해당한다. 다만 선행 연구(Nielson, 1994)에서는 개별 평가자가 내린 단일 점수는 신뢰성이 낮을 수 있으므로, 여러 평가자가 독립적으로 부여한 심각도 점수의 평균값을 활용하는 것이 실무적으로 유용하다고 보았다.[24] 그렇기에 본 연구 역시 이러한 방법론적 근거에 기반하여, 각 모델이 도출한 심각도 점수의 산술 평균(Mean)을 통해 모델 간의 정량적 평가 경향성을 비교 분석하였다.

3-3. LLM 사용성 평가 결과의 기초 분석

19) arXiv, LLM Stability: A Detailed Analysis with Some Surprises, (2026.05.12.) arxiv.org/abs/2408.04667

20) arXiv, Calibrating large language models with sample consistency, (2026.05.12.) arxiv.org/abs/2402.13904

21) Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J., 'Random effects structure for confirmatory hypothesis testing: Keep it maximal', Elsevier, Journal of Memory and Language, Vol.68, No.3, 2013.04, pp.255-278.

22) Bates, D., Mächler, M., Bolker, B., & Walker, S., 'Fitting linear mixed-effects models using lme4', Foundation for Open Access Statistics, Journal of Statistical Software, Vol.67, No.1, 2015.10, pp.1-48.

23) Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R., 'Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity', Elsevier, Computers and Education: Artificial Intelligence, Vol.9, 2025, p.100481.

24) Nielsen Norman Group, How to Rate the Severity of Usability Problems, (2026.06.01.) www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

본 절에서는 세 LLM의 휴리스틱 기반 UI 평가 결과에 대한 기술통계를 제시한다. 본 절은 기술통계 수준의 관찰적 비교이며, UI 수준 반복 구조와 집단 간 변동성을 통제한 통계적 검정은 3.4절의 혼합 효과 모형에서 수행한다.

3-3-1. 모델별 세션 평균 심각도

세 LLM(ChatGPT, Claude, Gemini) 간 세션 내(채팅별) 평균 심각도(sev_mean)를 비교한 결과, ChatGPT의 평균 심각도는 2.631로 가장 높게 나타났으며, Claude는 2.417, Gemini는 2.031의 값을 보였다. 이는 3가지 LLM 모델 간의 심각도 판단 기준에서 체계적인 차이가 존재함을 시사한다. 특히 ChatGPT는 상대적으로 높은 심각도를 부여하는 경향을 보였으며, Gemini는 세 모델 중 상대적으로 낮은 심각도 평가 경향을 보였다. 또한 ChatGPT와 Gemini 간 평균 차이(약 0.60)는 심각도 척도(1-4 범위) 를 고려할 때 무시하기 어려운 수준의 차이로 볼 수 있다. 이러한 결과는 모델별로 사용성 문제의 중요도를 해석하는 기준이나 내부 판단 임곗값이 상이할 가능성을 보여준다. 다만 본 절에서 제시한 결과는 기술통계에 기반한 평균 비교이며, UI 간 반복 측정 구조 및 집단 내 변동성을 통제하지 않은 상태의 관찰적 결과이다.

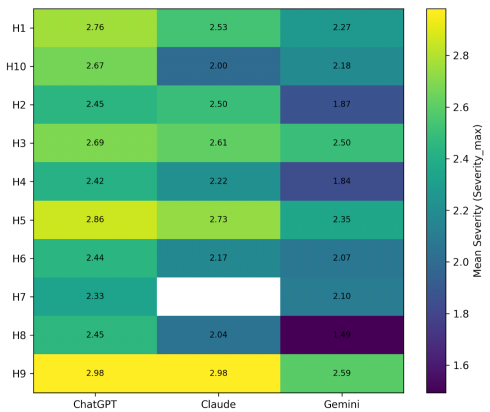
3-3-2. 모델별 세션 평균 문제 탐지 수 비교

세 LLM(ChatGPT, Claude, Gemini)의 세션 단위(채팅별) 평균 문제 탐지 수(issue_count)를 분석한 결과, 평균 문제 탐지 수가 ChatGPT 2.968, Claude 2.996, Gemini 2.972로 세 모델 간 차이는 0.028에 불과하였다. 이러한 결과는 세 모델이 전반적으로 유사한 양(amount)의 사용성 문제를 탐지하고 있음을 시사하며, 문제의 "수(quantity)" 측면에서는 모델 간 큰 차이가 관찰되지 않았다.

또한 표준편차를 비교하면 ChatGPT(SD=0.176)와 Gemini(SD=0.165)는 유사한 수준의 변동성을 보였지만, Claude는 상대적으로 낮은 표준편차(SD=0.063)를 나타내 세션 간 탐지 개수가 가장 일관되게 유지되었다. 그러나 이 결과를 해석할 때는 천장 효과(ceiling effect)의 가능성을 고려할 필요가 있다. 본 연구의 프롬프트는 모델이 보고할 수 있는 최대 문제 수를 3개로 제한하였는데, 전체 750개 세션 중 96.5%(724개)가 이 상한에 도달하였다(ChatGPT 94.4%, Claude 99.2%, Gemini 96.0%). 이는 모델들이 실제로 3개 이상의 문제를 탐지하였을 가능성이 있음에도 보고가 제한되었을 수 있음을 시사한다. 따라서 모델 간 문제 탐지 수의 차이가 유의하지 않다는 결과는 LLM의 실제 탐지 능력 차이가 없음을 의미하기보다, 프롬프트 설계상의 제약에 기인한 것으로 해석하는 것이 적절하다. 앞선 3.3.1절에서 확인된 심각도 평균 차이와 비교할 시, LLM 모델 간의 차이는 문제에 부여하는 심각도 판단 수준에서 더 뚜렷하게 나타나는 경향을 보여준다. 이는 모델 간 차이가 단순히 "더 많이 찾는가?"의 문제가 아닌, "어떤 문제를 얼마나 심각하게 해석하는가?"의 차이일 가능성을 시사한다.

3-3-3. 휴리스틱별 세션 평균 심각도

[그림 1]은 휴리스틱 항목(H1-H10)별 세션 평균 심각도(mean severity)를 모델에 따라 비교한 결과를 Heat-map으로 제시한다. 이는 전체 평균 수준을 넘어, LLM 모델 간 심각도 판단 차이가 특정 휴리스틱 항목에 집중되어 나타나는지를 탐색하기 위해 시행되었다. 전반적으로 ChatGPT는 대부분의 휴리스틱 항목에서 가장 높은 세션 평균 심각도를 보였으며, Gemini는 상대적으로 낮은 값을 나타내는 경향이 관찰되었다. 특히 H5, H8, H10 항목에서는 모델 간 평균 차이가 비교적 크게 나타났으며, 이는 해당 휴리스틱에 대한 문제 해석 및 중요도 판단 기준이 모델별로 상이할 가능성을 시사한다. 가령 H8 항목에서 ChatGPT(2.45)와 Claude(2.04)에 비해 Gemini(1.49)의 평균 심각도는 현저히 낮게 나타났다.



[그림 1] 휴리스틱 항목별 세션 평균 심각도 히트맵

반면 H3와 H9에서는 세 모델 간 평균값이 비교적

유사하게 나타나, 해당 항목에서는 모델 간 판단 경향 차이가 상대적으로 작을 것으로 유추된다. 또한 H7의 경우 Claude의 값이 관측되지 않았는데, 이는 해당 휴리스틱에 대해 Claude가 문제를 전혀 탐지하지 못한 결과에서 기인한 것으로 해석된다. 이러한 특이 현상은 특정 휴리스틱 항목에 대한 LLM 모델의 민감도 차이를 반영할 수 있으며, 이는 이후 분석에서 모델별 휴리스틱 적용 편향(heuristic sensitivity bias)의 가능성을 검토할 필요가 있음을 시사한다.

요약하자면, LLM 모델 간 평균 심각도 차이는 모든 휴리스틱 항목에서 균등하게 나타나는 것이 아닌, 특정 항목에 집중되는 경향을 보인다. 이는 LLM 모델별 차이가 휴리스틱 해석 구조의 차이에서 기인할 가능성을 암시한다.

3-4. 혼합 효과 모형에 기반한 가설 H1 검증

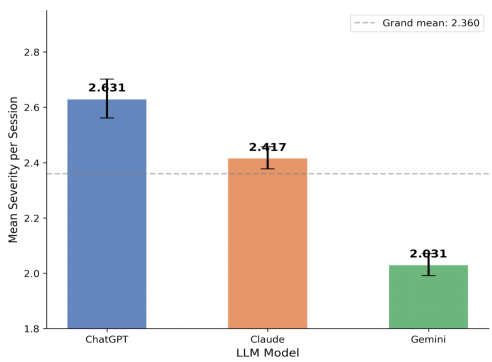
3-4-1. 분석 모형 설정

본 연구의 데이터는 UI, LLM 모델, 반복 평가로 구성된 계층적 구조를 가지므로, 분석에는 혼합 효과 모형(mixed-effects model)을 적용하였다. 휴리스틱 항목별 심각도 점수를 종속변수로 설정하고, Model을 고정 효과로 포함하였으며, UI_ID를 군집 요인으로 지정하여 UI 단위 반복을 반영하였다. 연구 데이터는 동일한 UI_ID에 대해 복수의 평가가 수행된 반복 측정 구조를 가지므로, 관측치 간 독립성이 엄격하게 충족된다고 보기 어렵다. 따라서 UI_ID를 랜덤효과(random effect)로 설정하여 UI 수준 변이를 반영하고, LLM 모델 유형(model)을 고정 효과(fixed effect)로 포함한 선형 혼합 효과 모형을 구성하였다. 모형은 다음과 같이 설정되었다. — 종속변수: sev_mean, 고정 효과: Model(ChatGPT를 기준 범주로 설정), 랜덤효과: UI_ID(random 절편 모형), 추정 방법: Restricted Maximum Likelihood(REML) 모형은 mean_severity ~ C(Model) 수식으로 설정하였으며, random intercept 구조를 적용하였다. 모든 통계 분석은 Python 환경에서 statsmodels 패키지를 활용하여 수행하였다. 다만 추정 과정에서 UI 수준 랜덤절편 분산은 0.054로 잔차 분산(0.053)과 유사한 수준이었으며, ICC = 0.50으로 나타나 전체 분산의 약 50%가 UI 간 차이에 기인함을 확인하였다. 이는 혼합 효과 모형 적용이 통계적으로 타당함을 뒷받침한다.

분석에 사용된 전체 관측치는 N = 750이며, UI_ID 기준 그룹 수는 50개이다. 각 그룹의 평균 관측치 수는 15로 확인되었다(mean group size = 15). 이는 각 UI가 세 모델에 대해 반복적으로 평가되었음을 의미하며, 혼합 효과 모형 적용의 통계적 타당성을 뒷받침한다. 이와 같은 모형 설정으로 LLM 모델 유형이 세션 평균 심각도에 미치는 효과를 UI 수준의 집단 변이를 통제한 상태에서 추정할 수 있다. 다음 절에서는 본 모형을 기반으로 LLM 모델 유형의 효과에 대한 구체적인 검증 결과를 제시하고자 한다.

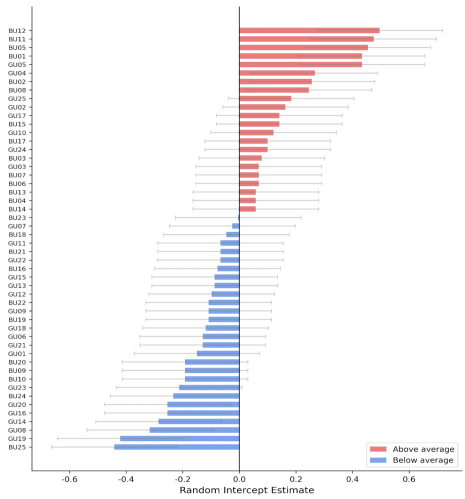
3-4-2. 모델 유형의 세션 평균 심각도에 대한 효과

혼합 효과 모형을 적용하여 모델 유형이 세션 평균 심각도(sev_mean)에 미치는 영향을 검증하였다. [그림 2]의 오차 막대는 혼합 효과 모형에서 도출된 95% 신뢰구간을 나타내며, 전체 평균(grand mean)은 2.360로 나타났다. ChatGPT를 기준 범주(reference category)로 설정한 결과, 절편(Intercept)은 2.631(SE = 0.036, $p < .001$)로 나타나 ChatGPT가 산출한 세션 평균 심각도의 기준 수준을 의미한다. [그림 3]의 붉은 막대는 평균 이상의 심각도를 보인 UI를, 파란 막대는 평균 이하의 심각도를 보인 UI를 나타낸다. 오차 막대는 랜덤효과 분포의 ±1 SD를 나타낸다. [표 10]의 LLM 모델별 고정 효과를 살펴보면, Claude는 ChatGPT 대비 세션 평균 심각도가 유의하게 낮았고(β = -0.214, SE = 0.021, z = -10.382, $p < .001$, 95% CI [-0.254, -0.174]), Gemini는 더욱 큰 폭으로 감소하였다(β = -0.600, SE = 0.021, z = -29.110, $p < .001$, 95% CI [-0.640, -0.560]). 실제 평균값은 ChatGPT 2.631, Claude 2.417, Gemini 2.031로 나타나 Gemini가 가장 낮고 ChatGPT가 가장 높은 심각도를 산출하는 경향을 보였다.



[그림 2] 혼합 효과 모형 기반 모델별 주변 평균 및 95% 신뢰구간

모형 추정 결과 분산 성분(variance component)은 0.054, 잔차 분산은 0.053이며, 급내상관계수(ICC) = .502로 전체 분산의 약 50%가 UI 간 차이에 기인하였다. 그러나 이러한 집단 구조를 통제한 이후에도 LLM 모델 유형의 효과는 통계적으로 매우 유의하게 유지되었다($p < .001$). 따라서 모델 유형이 심각도 판단에 영향을 미친다는 가설 H1은 일차적으로 지지가 된다.



[그림 3] UI별 랜덤 절편 추정값-Caterpillar plot

[표 10] 혼합 효과 모형 고정 효과 추정 결과

Predictor	β	SE	z	p	95% CI
Intercept (ChatGPT)	2.631	0.036	73.420	<.001	[2.561, 2.702]
Claude	-0.214	0.021	-10.382	<.001	[-0.254, -0.174]
Gemini	-0.600	0.021	-29.110	<.001	[-0.640, -0.560]

3-4-3. LLM 모델 간 효과 크기 분석

3.4.2.에서는 혼합 효과 모형을 통해 LLM 모델 유형이 세션 평균 심각도에 통계적으로 유의한 영향을 미침을 확인하였다. 그러나 통계적 유의성은 표본 크기에 영향을 받을 수 있으므로, 모델 간 차이의 실질적 크기를 평가하기 위하여 Cohen's d를 산출하였다. LLM 모델 간 평균 차이 및 효과 크기는 [표 11]과 같다. Mean Difference는 LLM 모델 간 평균 세션 심각도(sev_mean) 의 단순 차이이며, Cohen's d는 혼합 효과 모형의 총 분산(UI 랜덤효과 분산 0.054 + 잔차 분산 0.053) 의 제곱근(SD = 0.327)을 분모로 사용하여 다층 구조를 반영한 Cohen's d를 산출하였다. ChatGPT-Claude 간 d = 0.655(medium~large), ChatGPT-Gemini 간 d = 1.837(very large), Claude-Gemini 간 d = 1.182(large)로 나타났다.

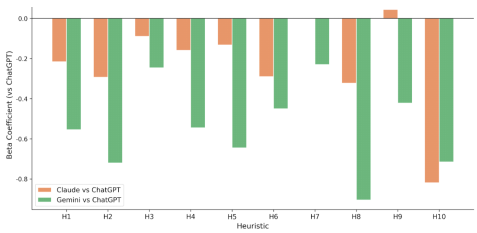
[표 11] LLM 모델 간 평균 차이 및 효과 크기

Comparison	Mean Difference	Cohen's d	Effect Size Interpretation
ChatGPT - Claude	0.214	0.655	Medium-Large
ChatGPT - Gemini	0.600	1.837	Very Large
Claude - Gemini	0.386	1.182	Large

세 모델 간 세션 평균 심각도 차이는 통계적 유의성을 넘어 효과 크기 측면에서도 중간 이상에서 매우 큰 수준에 이르며, 특히 Gemini는 다른 두 모델 대비 가장 큰 차이를 보였다. 이는 LLM 모델 유형이 심각도 판단에 실질적인 영향을 미친다는 가설 H1을 효과 크기 관점에서도 지지하며, 심각도 평가 기준의 강도 또는 판단 성향에서 구조적 차이가 존재할 가능성을 시사한다.

3-4-4. 휴리스틱 수준에서의 모델 효과 검증

3.4.2 그리고 3.4.3에서 LLM 모델 유형이 전체 세션 평균 심각도에 유의한 영향을 미침을 확인하였다. 그러나 이러한 효과가 모든 휴리스틱에서 동일하게 나타나는지를 검증하기 위하여, 각 휴리스틱(H1-H10)을 단위로 분리한 후 동일한 혼합 효과 모형을 적용하였다. 각 분석에서는 세션 평균 심각도(sev_mean)를 종속변수로 설정하고, Model을 고정 효과로 포함하였다. 또한 동일 UI에 대한 반복 평가 구조를 반영하기 위해 UI_ID를 군집 요인으로 포함한 혼합 효과 모형을 적용하였다. 일부 휴리스틱에서 특정 모델의 표본 수가 충분하지 않은 경우에는 OLS 모형으로 대체 분석을 수행하였다. 20회의 다중비교 문제를 통제하기 위해 Benjamini-Hochberg FDR 보정을 적용하였으며, 보정 후에도 모든 유의한 결과가 유지되었다(최소 p_{fdr} = .038, Claude H5). 휴리스틱별 모델 고정 효과 분석 결과는 [표 12]와 같다.



[그림 4] 휴리스틱별 모델 고정 효과 β 계수

[표 12] 휴리스틱별 모델 고정 효과 (기준: ChatGPT)

Heuristic	Claude β (95% CI)	p	p_fdr	Gemini β (95% CI)	p	p_fdr
H1	-0.216 (-0.302, -0.129)	< .001	0.000	-0.554 (-0.656, -0.451)	< .00	0.000
H2	-0.293 (-0.702, 0.116)	.161	.180	-0.720 (-1.019, -0.421)	< .00	0.000
H3	-0.089 (-0.202, 0.023)	.119	.141	-0.246 (-0.369, -0.122)	< .00	0.000
H4	-0.159 (-0.354, 0.037)	.112	.141	-0.545 (-0.697, -0.394)	< .00	0.000
H5	-0.132 (-0.249, -0.014)	.028	0.038	-0.645 (-0.802, -0.489)	< .00	0.000
H6	-0.290 (-0.395, -0.186)	< .001	0.000	-0.450 (-0.595, -0.305)	< .00	0.000
H7	—	—	NaN	-0.230 (-0.624, 0.163)	.24	0.263
H8	-0.323 (-0.431, -0.215)	< .001	0.000	-0.905 (-1.012, -0.798)	< .00	0.000
H9	0.045 (-0.128, 0.218)	.612	.612	-0.422 (-0.606, -0.237)	< .00	0.000
H10	-0.819 (-1.129, -0.510)	< .001	0.000	-0.715 (-0.925, -0.505)	< .00	0.000

분석 결과, Gemini는 H7을 제외한 모든 휴리스틱에서 ChatGPT 대비 세션 평균 심각도가 유의하게 낮은 것으로 나타났다($p < .001$). 특히 H1, H5, H6, H8 등에서 일관되게 음의 계수가 관찰되었으며, 이는 Gemini가 전반적으로 더 낮은 심각도 판단을 산출하는 경향이 휴리스틱 수준에서도 유지됨을 의미한다. 다만 H10의 경우 일부 모델의 표본 수가 6-22로 매우 제한적이었기에 (표 13), 추정치는 보고하되 해석은 보수적으로 접근할 필요가 있다. H7 또한 Claude의 관측치가 부재하여 모델 간 직접 비교가 불가능하였다. 이러한 결과는 3.4.2에서 확인된 전체 모델 효과가 특정 휴리스틱에 국한된 현상이 아니라 보다 구조적인 판단 경향임을 뒷받침한다. Claude의 경우 일부 휴리스틱(H1, H5, H6, H8 등) 항목에서 ChatGPT 대비 유의하게 낮은 심각도 값을 보였으나, H2, H3, H4, H7, H9 등의 항목에서는 통계적으로 유의한 차이가 관찰되지 않았다. 이는 Claude의 심각도 판단 차이가 모든 휴리스틱에서 동일하게 나타나는 것은 아니며, 특정 휴리스틱 유형에서 두드러지는 경향성을 시사한다. 또한 휴리스틱별 표본 분포를 확인한 결과, 일부 항목에서는 모델 간 표본 수 편차가 존재하였다. 예를 들어 H7에서는 특정 모델의 표본 수가 제한적(Claude의 표본이 존재하지 않음) 이었으며, 이에 따라 혼합 효과 모형이 수렴하지 않아 OLS 모형으로 대체 분석을 수행하였다. 이러한 표본 구조는 일부 휴리스틱에서 효과 추정의 안정성에 영향을 미칠 수 있으므로, 결과 해석 시 주의가 필요하다.

종합하면, 휴리스틱 수준 분석에서도 모델 유형은 일부 휴리스틱 항목에서 세션 평균 심각도 판단에 유의한 영향을 미치는 것으로 확인되었다. 특히 Gemini는 다수의 휴리스틱에서 일관되게 낮은 심각도 값을 산출하였으며, Claude는 특정 휴리스틱에서만 유의한 차이를 보였다. 이는 모델 간 판단 경향의 차이가 휴리스틱 특성에 따라 상이하게 나타날 수 있음을 의미한다. 따라서 H1a는 일부 휴리스틱 항목을 중심으로 부분적으로 지지가 된 것으로 해석할 수 있다.

3-4-5. 표본 분포 및 제한점

본 연구의 혼합 효과 모형 분석은 동일한 UI에 대해 다수의 LLM 모델 평가가 존재하는 반복 측정 구조를 고려하여, UI_ID를 군집 요인으로 포함하였다. 이를 통해 동일 UI에 속한 관측치 간의 잠재적 상관 구조를 모형 내에서 반영하였다. 그러나 휴리스틱 수준으로 분석 단위를 세분화할 경우, 휴리스틱별 표본 수는 상이하게 분포하였다. 휴리스틱별 모델 표본 분포는 [표 13]과 같다. 특히 일부 휴리스틱에서는 특정 모델의 평가가 존재하지 않거나 표본 수가 제한적인 경우가 확인되었다. 예를 들어 H7의 경우 Claude 모델의 평가 표본이 존재하지 않아 해당 항목에서는 혼합 효과 모형이 수렴하지 않았으며, 이에 따라 OLS 모형으로 대체 분석을 수행하였다. 이러한 표본 불균형은

효과 추정치의 분산을 증가시키거나 신뢰구간의 폭에 영향을 미칠 수 있다.

[표 13] 휴리스틱별 모델 표본 분포

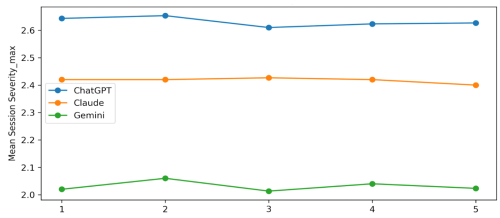
Heuristic	ChatGPT	Claude	Gemini
H1	171	171	96
H2	38	12	30
H3	103	95	66
H4	64	36	119
H5	69	90	37
H6	126	139	55
H7	6	0	136
H8	97	146	142
H9	44	53	37
H10	18	6	22

또한 휴리스틱별로 평균 그룹 크기(mean group size)가 상이하게 나타났으며, 최소 그룹 크기가 1인 경우도 존재하였다. 이는 특정 UI에서 일부 모델의 평가가 단일 관측치로 구성되었음을 의미하며, 랜덤효과 분산 추정의 안정성에 영향을 줄 가능성이 있다. 그런데도, 전체 모델 분석과 휴리스틱 수준 분석 모두에서 Gemini는 대부분의 항목에서 일관되게 낮은 세션 평균 심각도를 산출하는 경향을 보였으며, 이는 표본 구조의 제약에도 불구하고 비교적 안정적인 판단 패턴이 존재함을 시사한다. 반면 Claude는 일부 휴리스틱에서만 유의한 차이를 보였으며, 이는 표본 구조와 휴리스틱 특성의 영향을 동시에 반영한 결과일 가능성이 있다.

종합하자면 결과는 표본 분포의 불균형과 일부 휴리스틱에서의 제한적 표본 수라는 제약을 내포하고 있으나, 주요 효과는 다수의 휴리스틱에서 반복적으로 관찰되었다는 점에서 일정 수준의 구조적 일관성을 갖는 것으로 판단된다. 향후 연구에서는 휴리스틱별 균등 표본 설계 및 반복 평가 횟수를 확대하여 효과 추정의 안정성을 더욱 강화할 필요가 있을 것이다.

3-5. 실험 환경의 안정성 분석

본 절은 가설 H2를 검증하기 위해, 동일한 UI set가 동일한 프롬프트 조건에서 각 LLM 모델을 총 5회 반복 실행하여 세션 간 결과의 안정성을 분석하였다. 반복 실행(run)에 따른 세션 평균 심각도(mean severity per session)의 변화를 확인하기 위하여, 모델별 Run 평균값을 산출하고 이를 그림 5와 같이 시각화하였다. 그림 5에 제시된 바와 같이, 모든 모델에서 Run 1부터 Run 5까지 세션 평균 심각도 값의 변화 폭은 매우 제한적이었다. ChatGPT는 2.61-2.65 범위, Claude는 2.40-2.43 범위, Gemini는 2.01-2.06 범위 내에서 변동하였으며, 반복 실행에 따른 체계적 상승이나 하강 추세는 관찰되지 않았다. 이는 실험 조건에서 모델의 판단 결과가 실행 간 크게 변동하지 않았음을 시사한다. 더욱 정량적인 평가를 위해 Run 평균값의 표준편차(Run SD)와 변동계수(Run CV)를 산출하였으며, [표 14]에서는 반복 실행 간 세션 단위 변동성을 나타내었다.



[그림 5] 반복 실행 별 세션 평균 심각도 변화 추이

그 결과, ChatGPT의 Run SD는 0.017, Claude는 0.010, Gemini는 0.019로 나타났다. 변동계수(CV) 역시 각각 0.0065, 0.0042, 0.0093으로 모두 1% 미만 수준을 보였다. 이러한 수치는 반복 실행 간 변동성이 평균값 대비 극히 낮은 수준임을 의미한다.

[표 14] 반복 실행 간 세션 단위 변동성

Model	Run Mean	Run SD	Run CV	SD_across_runs_SevMean	SD_across_runs_issues
ChatGPT	2.631	0.017	0.0065	0.017095	0.017889
Claude	2.417	0.010	0.0042	0.010111	0.008944
Gemini	2.031	0.019	0.0093	0.018797	0.030332

특히 Gemini는 평균 심각도 수준이 가장 낮았으나, Run SD는 0.019로 다른 모델과 유사한 범위 내에 위치하였다. 이는 모델 간 절대적 심각도 차이는 존재하나, 동일 모델 내 반복 실행에 따른 판단 일관성은 안정적으로 유지되었음을 보여준다. 또한 모든 모델에서 Run CV가 0.01 이하로 나타난 점은 실험 환경 및 실행 조건이 반복 간 재현할 수 있게 유지되

었음을 뒷받침한다. 종합하면, 세션 단위(채팅별) 반복 실행 분석 결과, 본 연구의 실험 설계는 Run 간 높은 일관성을 확보한 것으로 판단할 수 있다. 추가적인 민감도 분석(sensitivity analysis)으로 세션 평균 심각도(sev_mean) 대신 세션 최대 심각도(sev_max)를 종속변수로 사용한 혼합 효과 모형을 실행한 결과, 모든 고정 효과의 방향과 유의성이 동일하게 유지되었다(Claude: β=-0.204, Gemini: β=-0.556, 모두 $p<.001$). 이는 본 연구의 주요 결과가 심각도 집계 방식과 관계없이 강건하게 유지됨을 의미한다. 따라서 3.4절에서 확인된 모델 간 세션 평균 심각도 차이는 일시적인 실행 변동이나 환경적 요인에 기인한 결과라기보다는, LLM 모델 유형에 따른 구조적 판단 경향의 차이를 반영하는 것으로 해석된다.

4. 결론

4-1. 연구 결과

본 연구는 LLM 기반 UX 휴리스틱 평가에서 모델 간 평가 경향 차이(inter-model evaluative tendency difference)가 존재하는지를 실증적으로 분석하고자 하였다. 동일한 UI와 동일 평가 기준(Nielsen의 휴리스틱)을 적용하고 반복 실행을 통제한 혼합 효과 모형 분석 결과, LLM 모델 유형은 전체 세션 평균 심각도에 통계적으로 유의한 영향을 미쳤다. UI 고유 난이도를 랜덤효과로 통제한 상태에서도 모델 간 평균 심각도 차이는 유의하였으며, 효과 크기 또한 중간 이상에서 매우 큰 수준에 이르렀다($d = 0.655~1.837$). 혼합 효과 모형의 ICC = 0.502로 전체 분산의 약 50%가 UI 간 차이에 기인하였고, 이는 UI 수준 랜덤 효과를 통제한 분석 설계의 적절성을 뒷받침한다. 또한 휴리스틱 항목별 분석에서도 일부 항목의 모델 간 차이가 다중비교 보정(FDR) 후에도 유의하게 유지되었으며, 동일 모델 내 반복 실행 변동성은 매우 낮았다(CV < 1%).

본 연구의 세 가설은 [표 15]에 제시된 바와 같이, H1과 H2는 완전히 지지가 되었으며 H1a는 부분적으로 지지가 되었다. H1a가 완전히 지지가 되지 않은 것은 모델 간 판단 차이가 모든 휴리스틱 항목에서 균일하게 나타나지 않고, 특정 항목(H5, H6, H8 등)에 집중되는 경향이 있음을 반영한다. 또한 H7에서 Claude의 평가 표본이 존재하지 않아 통계적 비교가 불가능했던 점은 결과 해석의 제한 요인으로 작용한다. 종합하면, 세 가설의 검증 결과는 LLM 모델 유형이 UX 사용성 평가 결과에 차이를 발생시킬 수 있음을 실증적으로 지지한다.

[표 15] 가설별 검증 결과 요약

가설	내용	검증 결과	근거
H1	LLM 모델 유형에 따라 휴리스틱 평가의 세션 평균 심각도에 통계적으로 유의미한 차이가 존재할 것이다.	지지	혼합 효과 모형 분석 결과, 모델 유형은 세션 평균 심각도에 유의한 영향을 미쳤으며($p < .001$), UI 수준 랜덤효과 통제 후에도 효과가 유지됨. 효과 크기 중간~매우 큰 수준($d = 0.655~1.837$)
H1a	H1의 모델 효과는 개별 휴리스틱 항목 수준에서도 관찰될 것이다.	부분 지지	Gemini는 H7을 제외한 대부분의 휴리스틱 항목에서 ChatGPT 대비 유의하게 낮은 심각도를 산출함. Claude는 H1, H5, H6, H8 등 일부 항목에서만 유의한 차이가 관찰됨. 모델 효과가 휴리스틱 특성에 따라 차별적으로 나타남
H2	동일 LLM 모델 내 반복 실행 간 심각도 판단은 높은 안정성을 유지할 것이다.	지지	모든 모델에서 Run CV < 1%(ChatGPT 0.65%, Claude 0.42%, Gemini 0.93%) 로, 반복 실행 간 판단 변동성이 극히 낮은 수준으로 확인됨

본 연구가 정의한 'LLM 모델 간 평가 경향 차이'는 정답의 오류나 환각이 아니라, 휴리스틱 기준을 적용하는 과정에서 나타나는 판단 임곗값(severity threshold)의 체계적 차이로 해석된다. 동일한 휴리스틱 정의와 심각도 척도를 제시했음에도 모델 간 평균 심각도 차이가 반복적으로 관찰되었다는 점은, LLM이 평가 기준을 해석하고 적용하는 방식이 모델별로 다를 가능성을 보여준다. 이러한 차이는 학습 데이터 구성, 내부 정책 등 다양한 구조적 요인의 복합적 결과일 수 있으나, 본 연구는 그 원인을 직접 규명하기보다 관찰 가능한 평가 경향의 존재를 확인하는 데 초점을 두었다. 다만 본 연구는 인간 전문가 기준선을 포함하지 않았으므로, 특정 모델의 판단이 더 정확하거나 우수하다고 해석하기보다는 모델 간 체계적 판단 차이의 존재를 확인한 연구로 이해할 필요가 있다.

4-2. UX 실무 및 연구 환경에 대한 함의

본 연구 결과는 LLM을 UX 사용성 평가 도구로 활

용하는 실무 및 연구 환경에 중요한 시사점을 제공한다. 다만 이하의 제안은 특정 모델의 판단이 더 정확하거나 실제 UX 품질을 더 잘 반영한다는 전제에 기반하지 않으며, 모델 간 판단 경향 차이가 존재한다는 본 연구의 관찰 결과를 실무적 맥락에서 해석한 것임을 밝힌다. 첫째, 모델 선택 자체가 평가 결과에 실질적인 영향을 미친다는 점을 인지해야 한다. 본 연구에서 ChatGPT(2.631)와 Gemini(2.031)의 세션 평균 심각도 차이는 0.600 점으로, 1-4점 척도에서 무시하기 어려운 수준이며 효과 크기 또한 매우 큰 수준(d = 1.837)으로 확인되었다. 이는 동일한 UI에 대해 어떤 LLM을 사용하느냐에 따라 사용성 문제의 심각도 판단이 구조적으로 달라질 수 있음을 의미한다. 따라서 UX 실무에서 LLM 기반 평가를 도입할 경우, 사용 모델을 명시하고 평가 결과를 해석할 때 해당 모델의 판단 경향을 함께 고려하는 것이 필수적이다.

둘째, 단일 모델 의존을 지양하고 복수 모델 기반 교차 검증 전략을 적용할 것을 권장한다. 구체적인 적용 방안으로는, 판단 경향이 상이한 복수의 모델로 동일 UI를 평가한 후 모델 간 심각도 점수의 평균값 또는 중앙값을 최종 심각도로 산출하는 방식을 고려할 수 있다. 본 연구의 결과를 기준으로 적용 시 ChatGPT와 Gemini는 심각도 판단 수준이 상이한 것으로 확인되었으므로, 두 모델을 병행 평가하여 판단 범위의 분포를 확인하는 방식을 고려할 수 있다. 또한 모델 간 심각도 차이가 특히 크게 나타난 H5, H8, H10 항목에서는 복수 모델 검증을 먼저 적용하는 것이 바람직하다.

셋째, LLM 평가 결과를 인간 전문가 평가와 병행하는 혼합 평가 체계(hybrid evaluation framework)의 도입을 권장한다. LLM은 반복 실행 간 변동계수(CV)가 1% 미만으로 매우 높은 판단 일관성을 보이므로, 대량의 UI를 빠르게 스크리닝하는 1차 평가 도구로서의 활용 가능성은 충분하다. 다만 본 연구에서 확인된 모델 간 판단 경향 차이를 고려할 때, 모델 간 심각도 판단이 불일치하거나 편차가 큰 항목을 인간 전문가가 우선적으로 검토하고, 모델 간 심각도 판단이 일치하지 않는 항목은 전문가 판단으로 최종 결정하는 2단계 평가 절차를 실무에 적용할 수 있다. 이러한 방식은 LLM의 비용·효율성 장점을 유지하면서도 모델 편향으로 인한 평가 왜곡을 최소화하는 현실적인 대안이 될 수 있다.

4-3. 연구의 한계

본 연구는 다음과 같은 한계를 가진다. 첫째, 본 연구는 LLM 간 평가 차이의 존재 여부를 검증하는데 초점을 두었으며, 특정 모델이 더 정확하거나 우수하다는 판단을 제공하지 않는다. 즉, 본 연구는 '차이의 존재'를 규명한 것이지 '정확성의 우열'을 판정한 것은 아니다.

둘째, 본 연구는 인간 UX 전문가 평가를 기준선(human baseline)으로 포함하지 않았다. 이는 본 연구의 목적이 "LLM 모델 간 차이의 존재"를 검증하는데 있었기 때문이며, "어느 모델의 판단이 인간 전문가 평가에 더 근접하는가?"는 본 연구의 범위를 벗어난다. 그러나 이 점은 동시에 본 연구의 해석상 한계를 구성한다. 인간 기준선이 부재한 상태에서는 ChatGPT의 높은 심각도 판단이 과대평가인지, Gemini의 낮은 심각도 판단이 과소평가인지 단정할 수 없으며, 본 연구는 그러한 방향성(directionality) 판단을 의도적으로 유보하였다. 향후 연구에서 인간 전문가 평가 데이터를 기준선으로 포함할 경우, 본 연구가 관찰한 모델 간 차이를 정확성의 관점에서 재해석할 수 있을 것이다. 따라서 본 연구는 모델 간 평가 경향 차이를 분석한 것이며, 어떤 모델의 판단이 더 정확한지에 대한 검증은 인간 전문가 기준선을 포함한 향후 연구 과제로 명확히 남겨둔다.

셋째, 일부 휴리스틱 항목에서는 표본 수가 제한적이었으며, 특정 모델이 해당 항목을 전혀 지적하지 않은 경우(예: Claude의 H7), 통계적 해석에 제약이 존재한다. 이러한 불균형은 항목별 결과 해석 시 신중한 접근을 요구한다. 넷째, 본 연구는 이미지 기반 UI와 텍스트 시나리오를 중심으로 평가를 수행하였으며, 실제 인터랙션 기반 사용성 테스트와는 맥락적 차이가 존재할 수 있다. 또한 프롬프트를 투입함에 있어 LLM 모델이 자연어를 해석하면서 생기는 차이가 발생할 수 있다. 추가로 프롬프트 설계상 세션당 최대 3개의 문제만 보고하도록 제한하였으며, 실제 분석 결과 전체 세션의 96.5%가 이 상한에 도달하였다. 향후 연구에서는 보고 개수 제한을 완화하거나 제거하여 모델의 실제 탐지 능력을 더욱 정확하게 평가할 필요가 있다. 또한 코딩 작업이 단일 연구자에 의해 수행되었으며 코더 간 신뢰도(inter-coder reliability) 검증이 이루어지지 않았다는 점은 본 연구의 방법론적 한계로 남는다. 향후 연구에서는 복수의 코더를 활용한 신뢰도 검증 절차를 도입하여, 코딩의 타당성을 강화할 필요가 있다. 마지막으로, 평가 편향의 발

생 원인을 실험적으로 규명하지 않았다는 점에서 설명적 한계를 가진다.

4-4. 향후 연구 방향

향후 연구에서는 다음과 같은 확장이 가능하다. 첫째, 본 연구에서 확인한 모델 간 평가 경향 차이가 실제 UX 품질 판단에서 어떠한 의미를 갖는지 검증하기 위해, 인간 전문가 평가를 기준선으로 포함한 후속 연구가 필요하다. 이를 통해 각 LLM 모델의 심각도 판단이 전문가 판단에 얼마나 근접하는지, 그리고 모델 간 차이가 정확도 차이로 해석될 수 있는지를 실증적으로 규명할 수 있을 것이다. 둘째, 다양한 도메인으로 범위를 확장하여 평가 경향 차이의 일반화 가능성을 검증할 필요가 있다. 셋째, 프롬프트 구조 변화 혹은 LLM의 역할 부여 전략이 모델 간 판단 경향 차이에 미치는 영향을 실험적으로 분석함으로써, 경향 차이 조절 가능성을 모색할 수 있다. 넷째, 다중 LLM 기반 평가 프레임워크를 설계하여 모델 간 차이를 상쇄하거나 보완하는 방법을 연구할 수 있다. 이를 통해 LLM 기반 UX 사용성 평가의 구조적 한계를 체계적으로 이해하고, 신뢰성을 향상하는 방향으로 연구를 확장할 수 있을 것이다.

참고문헌

1. 김민채, 박상희, '디자인 프로세스 수행 효율성 향상을 위한 LLM 프롬프트 엔지니어링 기법 연구-디자인씽킹을 중심으로-', 한국브랜드디자인학회, 브랜드디자인학연구, 2025
2. 김유진, 강조은, 김한샘, '언어모델의 편향 개선을 위한 프롬프트 엔지니어링 연구: ChatGPT를 활용한 정치인 감성분석 말뭉치를 중심으로', 한국코퍼스언어학회, Corpus Linguistics Research, 2023
3. 김현웅, 안현철, '대규모 언어 모델(LLM) 응답에 내재된 미묘한 성차별에 대한 연구', 한국빅데이터학회, 한국빅데이터학회지, 2025
4. 방준성, 이병탁, 박판근, '대규모 언어 모델을 사용하는 인공지능 기반 대화형 챗봇의 편향성 평가 프레임워크 개발 방법', 한국방송•미디어공학회, 방송공학회 논문지, 2023
5. 이서후, 김승인, 'LLM기반 AI 서비스 사용자경험 비교연구: ChatGPT와 Clova X를 중심으로', 한국상품문화디자인학회, 상품문화디자인학연구, 2023
6. 이주은, 배호, '합성 데이터를 활용한 Large Language Model의 인종 및 성별 교차 편향 측정 벤치마크', 한국정보처리학회, 정보처리학회 논문지, 2025
7. Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J., 'Random effects structure for confirmatory hypothesis testing: Keep it maximal', Elsevier, Journal of Memory and Language, 2013
8. Bates, D., Mächler, M., Bolker, B., & Walker, S., 'Fitting linear mixed-effects models using lme4', Foundation for Open Access Statistics, Journal of Statistical Software, 2015
9. Campos, Luis F. G., Marques, Leonardo C. & Nakamura, Walter T., 'Will AI also replace inspectors? Investigating the potential of generative AIs in usability inspection', ACM, Proceedings of the Brazilian Symposium on Software Quality(SBQS '25), 2025
10. Liu, Yang, Iter, Dan, Xu, Yichong, Wang, Shuohang, Xu, Ruochen & Zhu, Chenguang, 'G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment', Association for Computational Linguistics, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing(EMNLP 2023), 2023
11. Nielsen, Jakob & Molich, Rolf, 'Heuristic Evaluation of User Interfaces', ACM, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems(CHI '90), 1990

12. Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R., 'Evaluating Large Language Models as Raters in Large-Scale Writing Assessments: A Psychometric Framework for Reliability and Validity', Elsevier, Computers and Education: Artificial Intelligence, 2025
13. arxiv.org
14. toss.tech
15. www.nngroup.com