JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 12, pp. 3789-3797, Dec. 2024

Check for updates

정보 검색을 위한 AI 서비스의 사용자 경험 비교 연구: Naver CUE:, Bing Copilot을 중심으로

박 창 준[1] · 이 정 진[2*]
[1]숭실대학교 글로벌미디어학부 학사과정
[2]숭실대학교 글로벌미디어학부 조교수

Comparative Study on User Experience of AI-Based Information Retrieval Services: Focusing on Naver Cue and Bing Copilot

Changjun Park[1] · Jungjin Lee[2*]

[1]Bachelor's Degree Program, Department of Global Media, Soongsil University, Seoul 06978, Korea
[2]Assistant Professor, Department of Global Media, Soongsil University, Seoul 06978, Korea

[요 약]

본 연구는 Naver Cue:와 Bing Copilot을 중심으로, 정보 검색 분야에서 LLM 기반 생성형 AI 서비스가 어떤 장단점이 있는지 분석한다. 사용성, 유용성, 신뢰성, 매력성, 윤리성의 사용자 경험 평가 요소를 기반으로 두 서비스를 비교한 결과, 매력성에서 유의미한 차이가 있음을 확인할 수 있었다. 사용자 경험 인터뷰 결과, Naver Cue:는 한국에 최적화된 서비스라는 점과 사용자가 원하는 조건에 충실한 정보 제공으로 긍정적인 평가를 받았으나, 일부 원치 않는 정보가 제공되는 문제가 있었다. Bing Copilot은 빠른 검색 속도와 정보 정리 능력에서 긍정적인 평가를 받았으나, 사용자가 요구한 조건을 누락하는 문제와 UI 디자인 개선이 필요하다는 지적이 있었다. 이러한 분석을 바탕으로 개선 방안을 제시하며, AI 서비스가 문제점을 극복하고 대중화될 수 있기를 기대한다.

[Abstract]

This study analyzes the advantages and disadvantages of generative artificial intelligence (AI) services based on large language model (LLM) in the field of information retrieval, focusing on Naver Cue and Bing Copilot. Based on user experience evaluation criteria, such as usable, useful, credible, desirable, and moral, a significant difference was observed in the attractiveness of these two services. According to user experience interviews, Naver Cue received positive evaluations for feeling like a service optimized for Korea and providing information that closely matches users' desired conditions despite issues with some unwanted information being provided. Bing Copilot was praised for its fast search speed and information organization capabilities but was criticized for missing user-specified conditions and requiring improvements in user interface (UI) design. Based on this analysis, the study suggests improvement measures, hoping that these AI services can overcome the aforementioned problems and become more widely adopted.

색인어 : 인공지능, 생성형 AI, 거대언어모델, 정보 검색 서비스, 사용자 경험
Keyword : Artificial Intelligence, Generative AI, Large Language Model, Information Retrieval Service, User Experience

http://dx.doi.org/10.9728/dcs.2024.25.12.3789

Received 31 August 2024; **Revised** 07 October 2024
Accepted 16 October 2024

***Corresponding Author; Jungjin Lee**

Tel: +82-2-820-0918
E-mail: jungjinlee@ssu.ac.kr

Ⅰ. 서 론

OpenAI의 GPT-3.5를 기반으로 한 ChatGPT의 등장 이후, LLM(Large Language Model)을 활용한 AI(Artificial Intelligence) 서비스가 급격히 대중화되었다. 과학기술정보통신부가 2024년 3월에 발표한 '2023 인터넷 이용 실태조사'에 따르면, 2021년부터 3년간 인공지능 경험률은 지속적으로 증가하여, 2023년에는 50.8%에 도달해 응답자 2명 중 1명 이상이 인공지능 서비스를 경험한 것으로 나타났다. 인공지능 사용 경험은 주거 편의 분야, 특히 가전과 관련된 분야에서 20.8%로 가장 높았으며, 교육 분야에서는 가장 빠른 증가 추세를 보였다. 이처럼 AI는 다양한 일상생활 영역에 확산되어 우리의 삶을 변화시키고 있다[1].

AI는 정보 검색 분야(IR)에서도 새로운 검색 경험을 제공하며 기존 방식과는 다른 접근법을 제시하고 있다. 전통적인 검색은 사용자가 키워드 텍스트를 입력하면 검색 엔진이 해당 텍스트와 일치하는 정보를 기반으로 결과를 도출하는 방식이어서, 키워드에 지나치게 의존하고 맥락을 충분히 감지하지 못하는 한계가 있었다. 그러나 생성형 AI를 활용하면 사용자와의 문답을 통해 AI가 자동으로 정보를 요약하고 정리하여, 사용자에게 맞춤형 정보를 제공할 수 있다[2].

이러한 AI 기반의 정보 검색 혁신은 여러 검색 플랫폼에서 활발히 진행되고 있다. 구글(Google)은 제미나이(Geminai) 모델을 구글 검색 시스템에 결합한 'AI 개요 기능'을 통해 검색 결과를 생성형 AI로 요약하여 제공하고 있다[3]. 마이크로소프트(Microsoft)는 GPT-4 모델을 바탕으로 AI 어시스턴트인 '코파일럿(Copilot)'을 제작하여 빙(Bing)에서 지능적인 검색 경험을 제공한다[4]. 한국에서는 네이버(Naver)가 하이퍼클로바X(Hyper Clova X)를 기반으로 검색 서비스에 특화된 '네이버 큐(Naver Cue:)'를 개발하여, 정보 검색 경험에 AI를 접목하고 있다[5].

그러나 AI의 기술적 발전에도 불구하고, 정보 검색 분야에서 LLM 기반 생성형 AI가 해결해야 할 몇 가지 과제가 남아 있다. 첫 번째 과제는 '환각(Hallucination)' 현상으로, 이는 LLM이 문맥을 이해한 것처럼 보이나 실제로는 사실에 기반하지 않은 틀린 답변을 그럴듯하게 생성하는 현상을 말한다. 이에 따라 사용자는 잘못된 정보를 바탕으로 의사결정을 내릴 위험이 있으며, 비즈니스나 학문 분야와 같이 정확성이 중요한 상황에서는 환각 현상이 경제적 손실이나 사고로 이어질 가능성이 있다[6]. 두 번째 과제는, 아직 초기 시장이기에 캐즘(Chasm)으로 인해 앞으로 대중화에 어려움을 겪을 수 있다. 캐즘이란 초기 시장의 혁신 기술이 성장하는 과정에서 대중에게 널리 수용되는 주류 시장으로 도달하기 전에 매출이 급감하거나 정체현상이 발생하는 침체기를 의미한다. 초기 혁신 기술을 받아들이는 선도자들은 신기술 도입에 적극적이지만, 주류 시장의 대다수는 실용성을 중시하는 실용주의자들이기 때문에, 기술이 실용적인 문제를 해결하지 못하면 대중의 수요를 이끌어내지 못하고 결국 사라질 가능성이 있다[7].

본 연구는 AI를 활용한 정보 검색이 사용자가 정보를 효과적으로 검색하는 데 어떠한 영향을 미치는지 실증하고자 한다. 특히, 현재 출시된 여러 LLM 기반 AI 정보 검색 서비스를 비교 분석을 통해 장단점을 도출하여 서비스 마다의 장점을 참고해 각 서비스의 한계점을 극복할 수 있는 개선 방안을 도출한다. 이를 통해 AI 서비스가 캐즘을 극복하고 보다 널리 채택될 수 있는 방향성을 모색하고자 한다.

Ⅱ. 이론적 배경

2-1 사용자 경험 선행 연구 방식

AI의 발전은 UX(User Experience)에 혁신적인 변화를 불러왔다. UX는 사용자가 제품 또는 서비스를 이용하면서 느끼는 신체적, 감정적, 정신적 반응을 포함한 총체적인 경험을 의미한다[8]. 서비스의 UX를 발전시키기 위해서 사용자 경험을 평가하기 위한 기존의 평가 요소들이 있다. 제이콥 닐슨(Jakob Nielson)은 사용성을 평가하는 요소로 만족성(Satisfaction), 오류(Errors), 기억성(Memorablility), 학습성(Learnablility), 효율성(Efficiency)을 제안했다[9]. ISO 9241-11에서는 필요와 목적에 맞는 효과성(Efficiency), 감성적인 만족도(Satisfaction), 사용의 효율성(Efficiency)을 평가 요소로 선정했다[10]. 서비스의 UX를 평가하는 선행 연구에서 기존의 평가 요소들에 따라 연구를 진행해 왔으나, AI 기반 서비스를 평가하기 위해서는 다른 평가 요소가 필요하다. 그 이유는 AI는 안전과 관련된 불확실성을 가지고 있기 때문에 도덕성에 대한 평가가 필요한데 기존 평가 요소는 이러한 특성을 반영하지 못한다. 그렇기에 AI 서비스의 사용자 경험에 대한 선행 연구는 LLM 기반의 AI 서비스 특성에 맞는 사용자 경험 평가 요소를 정립한다[11]. 해당 연구에서는 도덕성과 더불어 개인화된 경험에 대한 중요성을 반영하여 평가한다. AI 기반 UX의 주요 이점 중 하나는 개별 사용자에게 고도로 개인화된 경험을 제공할 수 있다는 점이 있기에 이를 평가할 수 있는 요소가 중요하다. AI는 사용자 데이터를 분석하여 사용자의 행동과 선호도에 따라 개인화된 추천을 제공한다. 예를 들어, 검색 기록, 구매 기록, 소셜 미디어 활동 등의 데이터를 AI가 분석하여, 사용자에게 최적화된 추천을 제공한다. 이로 인해 사용자는 자신에게 가장 관련성이 높은 콘텐츠와 기능을 제공받아 더 만족스러운 사용자 경험을 누릴 수 있다[12].

2-2 사용자 경험 평가 요소

LLM 기반 AI 서비스의 사용자 경험을 평가하기 위해서 피터 모빌(Peter Morvile)의 허니컴 모델(Honeycomb)과 Chat GPT 사용자 경험 디자인 요소를 참고했다. 피터 모빌은

2004년에 '검색 2.0: 발견의 진화 ambient findability'에서 훌륭한 UX를 정의하기 위한 7가지 평가 요소로 분석하는 허니컴 모델을 제안한다[13]. 그림 1과 같이 허니컴 모델에는 사용성(Usable), 유용성(Useful), 매력성(Desirable), 가치성(Valuable), 접근성(Accessable), 검색성(Findable), 신뢰성(Credible)으로 7가지 평가 요소가 있다.

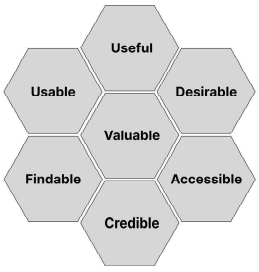

그림 1. 피터 모빌의 허니컴 모델
Fig. 1. Peter Morville's user experience honeycomb

사용성은 사용자가 서비스를 이용하는 데 어려움이 없는지를 평가하는 요소이다. 유용성은 서비스가 기능적으로 가치가 있으며 사용자의 요구를 충족시키는지를 평가한다. 매력성은 사용자가 서비스를 이용하면서 감정적으로 만족하거나 즐거움을 느끼는지를 평가한다. 접근성은 사용자가 서비스에 쉽게 접근할 수 있는지를 평가하고, 검색성은 서비스 내에서 콘텐츠와 기능을 쉽게 찾을 수 있는지를 측정한다. 마지막으로, 신뢰성은 서비스가 사용자를 속이지 않고 믿을 수 있는지를 평가하는 요소이다[14].

허니컴 모델 외에도 선행 연구에서 제시된 ChatGPT 사용자 경험 디자인 설문 항목을 참고하였다. 이 연구를 참고한 이유는 AI 서비스에 특화된 설문 항목을 포함하고 있기 때문이다. 설문은 유용성, 신뢰성, 성능, 공정성, 사회적 책임, 사용성을 중심으로 구성되었으며, 리커트 척도를 사용하여 평가되었다. 설문 항목의 구체적인 내용은 다음과 같다. 유용성은 ChatGPT가 업무나 일상생활에서 발생하는 문제를 해결하거나 목적을 달성하는 데 도움이 되는지를 평가하며, 신뢰성은 ChatGPT가 생성한 결과의 내용과 맥락이 일관성을 유지하여 사용자가 원하는 정보를 제공하는지를 평가한다. 성능 항목은 ChatGPT의 처리 속도와 답변 정확도를 측정하고, 공정성은 ChatGPT가 나이, 성별, 지역 등과 같은 인적 특성에 대한 편견 없이 결과를 생성하는지를 평가한다. 사회적 책임 항목은 ChatGPT가 사회적 가치를 위협하거나 유해한 질문을 거부하는지를 다루며, 사용성은 ChatGPT가 사용이 간편하며 원하는 정보를 검색하는 데 효과적인지를 평가하는 내용으로 구성되었다[15].

본 연구는 피터 모빌의 허니컴 모형과 ChatGPT 사용자 경험 디자인 설문 항목을 기반으로 평가 요소를 구성하여 표 1과 같이 정리하였다. 평가 모델에서 유사한 평가 요소를 합치고, AI를 평가하는 데 필요한 평가 요소를 보충하여 새롭게 평가 요소를 도출했다. 총 5가지 평가 요소로 사용성(Usable), 유용성(Useful), 신뢰성(Credible), 매력성(Desirable), 도덕성(Moral)으로 구성하였다.

표 1. 사용자 경험 평가 요소와 정의
Table 1. User experience evaluation factors and definitions

Factors	Definition
Usable	The design must be easy to use and navigate.
Useful	The Content and functionality must be valuable and fulfill a user need.
Credible	The content and design must convey trust and authenticity.
Desirable	The Design should evoke emotion and appreciation, making the experience enjoyable.
Moral	The Service must reject conversations that threaten social values or are harmful.

Ⅲ. 연구 방법

3-1 연구 대상 서비스 선정

정보 검색 분야에 LLM을 기반하는 AI 검색 서비스 중 네이버에서 개발한 Naver CUE:와 마이크로소프트의 Bing Copilot을 중점으로 분석했다. 구글의 제미나이 AI 개요 서비스는 연구를 실행한 2024년 5월 기준 미국에서만 서비스하기 때문에 연구에 포함할 수 없었다[16].

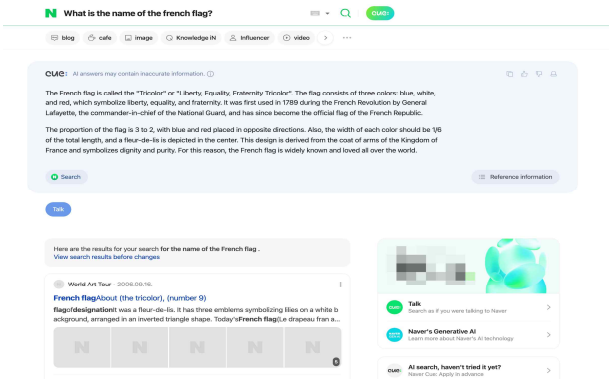

그림 2. Naver CUE:의 검색 결과 화면
Fig. 2. Retrieval results screen of Naver CUE:

Naver CUE:와 Bing Copilot은 Chat GPT와 유사하게 자연어로 질문할 수 있는 AI 기반 서비스이다. 그러나 이 두 서비스는 정보 검색에 특화된 UX를 제공하는 점에서 차별화된다. 그림 2, 그림 3에서 볼 수 있듯이, 검색창 옆에 바로 접근할 수 있는 아이콘을 제공하여 사용자가 검색 중에 쉽게 AI를 활용할 수 있도록 한다. 또한, 인공지능은 검색 결과를 검색창 하단에 정리하여 보여준다. 이는 사용자가 별도로 페이지를 이동할 필요 없이, 기존의 검색 과정에서 자연스럽게 AI의 도움을 받을 수 있는 환경을 조성한다.

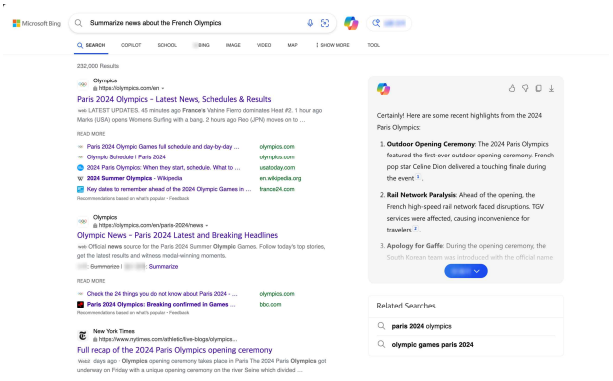

그림 3. Microsoft Bing Copilot의 검색 결과 화면
Fig. 3. Retrieval results screen of Microsoft Bing Copilot

이렇듯 Naver CUE:와 Bing Copilot은 정보 검색 분야에 특화된 특성을 가져 연구 대상 서비스로 선정했다. 사용자가 두 서비스를 이용하는 모습을 관찰하고, 설문조사와 인터뷰를 진행한다.

3-2 연구 진행 방식

연구의 진행 방식은 그림 4와 같다. 연구를 진행하기에 앞서서 연구 참여 대상자에게 준비 사항을 미리 공지하였다. Naver CUE:는 2024년 기준 베타 서비스로, 사전 신청자만 사용할 수 있었기 때문에 참여자들이 미리 신청할 수 있도록 했다. Bing Copilot의 경우, Microsoft Edge를 설치하고 Bing 내에서만 사용할 수 있어, 연구 참여자들에게 해당 프로그램을 설치하도록 안내했다. 연구 시작에 앞서, 각 서비스에 익숙해질 수 있도록 5분간 사용 시간을 제공한다. 이는 연구 과정이 원활하게 진행되도록 하기 위함이다. 이후 연구 목적과 과정에 대해 안내한 후, 주제에 따라 각 참여자가 태스크를 수행한다.

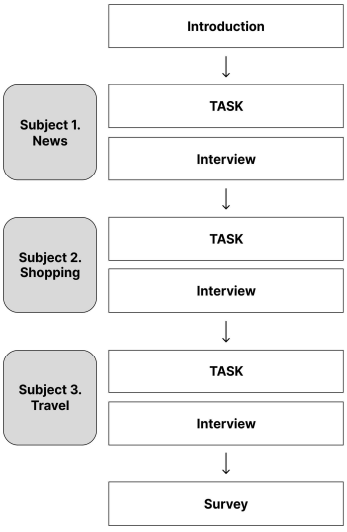

그림 4. 연구 절차
Fig. 4. Study process

연구는 평균적으로 1시간 동안 진행되었으며, 표 2와 같이 총 3개의 주제에 대한 태스크를 수행한다. 연구 태스크의 각 주제는 PC에서 사용자가 검색하는 주제에 대한 논문을 참고하였다. 해당 연구에 따르면, PC에서 검색하는 주제는 학업/연구가 가장 많이 검색하는 주제였으며, 이어서 쇼핑, 지역/여행 순으로 나타났다[17]. 이를 참고하여 주제를 뉴스, 쇼핑, 여행으로 선정했다. 학업/연구에 대한 주제를 뉴스로 선정한 이유는 뉴스에서 사용자가 모르는 내용 혹은 용어 설명을 AI가 얼마나 친절하게 설명하는지를 평가하기 때문에 모르는 개념을 학업 하는 과정을 경험하는 것으로 판단했기 때문이다. 주제마다 2개의 태스크를 이어서 진행했다. 한 주제의 태스크가 끝나면 인터뷰를 진행하여 연구 참여자가 서비스를 사용하면서 어떤 감정을 느꼈는지, 어떤 불편함을 느꼈는지 파악했다. 모든 태스크가 완료되면 Naver CUE:와 Bing Copilot 각각 평가 요소에 대한 리커트 척도 설문을 진행했다.

표 2. 주제 별 연구 태스크
Table 2. Research tasks by topic

Subject	Task
News	1. Please search for today's news issues. 2. Please ask additional questions for clarification if the explanation is unclear.
Shopping	1. Please describe the recipient of the gift and recommend a gift. 2. Please select one of the recommended gift types to find a product you would like to purchase.
Travel	1. Please explore the travel destination you want to go to. 2. Please receive recommendations for restaurants in the travel area and select the final location to go.

표 3. 연구 표본의 인구 통계학적 특성
Table 3. Demographic characteristics of the study sample

Demographic Factors		Frequency(N)	Proportion(%)
Gender	Male	12	48
	Female	13	52
Age	20-24	4	16
	25-29	17	68
	30-34	4	16
Job	College Student	7	28
	Unemployed	5	20
	Company Employee	13	52
Sum		25	100

3-3 연구 표본

연구 표본으로 20대 이상 성인 남녀 총 25명을 연구했다. 빈도 분석 결과는 표 3과 같다. 성별은 남성 12명(48%)과 여성 13(52%) 비율이다. 나이는 20-24세는 4명(16%), 25-29

세는 17명(68%), 30-34세는 4명(4%)이다. 직업은 대학생 7명(7%), 무직 5명(5%), 직장인 13명(52%)이다.

Ⅳ. 분석 결과

4-1 분석 방법

사용성, 유용성, 신뢰성, 매력성, 윤리성을 리커트척도 분석을 진행했다. 본 연구의 자료처리는 SPSS 24.0 프로그램을 사용했으며, 가설검정의 유의수준은 0.05로 설정했다. 연구의 절차는 다음과 같다.

첫째, 변수의 정규성을 확인하기 위해 Shapiro-Wilk 검정을 수행했다. 둘째, 변수의 기술통계와 Naver CUE:와 Bing Copilot의 사용성, 유용성, 신뢰성, 매력성, 윤리성의 차이 검정을 시행했다. 차이 검정 방법은 앞서 정규성 검정 결과에 따라 비모수적 방법인 Wilcoxon 부호순위 검정을 수행했다.

4-2 정규성 검정

Naver CUE:와 Bing Copilot의 사용성, 유용성, 신뢰성, 매력성, 윤리성의 정규성 검정 결과는 다음 표 4와 같다. p-value가 0.05보다 작은 값은 정규성을 만족하지 않는다. 따라서 Shapiro-Wilk 검정 결과, Naver CUE:에서는 윤리성이, MS Bing에서는 유용성, 신뢰성, 윤리성이 정규성을 만족하지 않은 것으로 나타났다. 따라서 본 분석의 차이 검정은 정규성 가정을 만족하지 않은 것을 고려하여 비모수적 방법으로 수행했다.

표 4. 정규성 검정 결과
Table 4. Normality test results

Service	Variable	Shapiro-Wilk	df	p-value
Naver CUE:	Usable	.921	25	.053
	Useful	.931	25	.092
	Credible	.923	25	.059
	Desirable	.962	25	.456
	Moral	.875	25	.005**
Bing Copilot	Usable	.965	25	.516
	Useful	.914	25	.037*
	Credible	.911	25	.032*
	Desirable	.937	25	.125
	Moral	.880	25	.007**

*p<.05, **p<.01, ***p<.001

4-3 Naver CUE:와 Bing Copilot의 차이 검정 결과

본 연구에서는 동일한 실험자를 대상으로 Naver CUE:와 Bing Copilot의 사용성, 유용성, 신뢰성, 매력성, 윤리성을 측정했다. 차이 검정은 비모수적인 방법인 Wilcoxon 부호순위 검정으로 수행했으며 결과는 표 5와 같다.

차이 검정 결과, 사용성, 유용성, 신뢰성, 윤리성에서는 Naver Cue:와 Bing Copilot의 유의미한 차이가 없었다. 매력성의 경우 Naver Cue:의 평균은 3.55, Bing Copilot의 평균은 3.01로 Naver Cue:가 높았으며, 차이 검정 결과 유의확률은 0.05보다 작은 것으로 나타나 차이가 유의미함을 확인했다(Z=-2.337, p<.05).

표 5. Wilcoxon 부호순위 검정 결과
Table 5. Wilcoxon signed rank test results

Variable	Naver Cue:		Bing Copilot		Z	p-value
	Avg	SD	Avg	SD		
Usable	3.85	.55	3.59	.68	-1.671	.095
Useful	3.08	.80	3.30	.75	-.785	.433
Credible	3.68	.56	3.32	.69	-1.911	.056
Desirable	3.55	.67	3.01	.68	-2.337	.019*
Moral	4.12	.70	4.14	.74	-.395	.693

*p<.05, **p<.01, ***p<.001

표 6. '사용성' 변수에 대한 인터뷰 결과
Table 6. Interview results for variable 'Usable'

Variable	Service	Positive Comments
Usable	Naver CUE:	• The functions are intuitive. • Because it is a platform optimized for Korean, the information is easy to read. • It is kind because it shows the steps to process information.
	Bing Copilot	• The functions are intuitive. • Structured answers allow for smooth delivery of key content. • The speed of providing answers is fast.
	Service	Negative Comments
	Naver CUE:	• There is a lot of text, so it is not intuitive to understand the content. • In some cases, there is insufficient information and additional questions may be needed.
	Bing Copilot	• I feel that it does not provide optimal service to Koreans. • The spacing between words is narrow, and the font is not easy to read.

4-4 사용자 경험 인터뷰 결과

인터뷰 결과, 실험자들은 사용성에 대해서 Naver CUE:와 Bing Copilot 모두 대체로 직관적인 기능을 가졌다고 느꼈다. Naver CUE:는 한국어에 최적화되어 내용을 이해하는데 용이했고, 검색 결과를 도출하는 과정을 보여주어 친절하다는 인상을 주었다. 그러나 많은 글자로 인한 가독성 저하, 여러 번 질문해야 원하는 답을 얻을 수 있다는 불편함이 지적되었

다. Bing Copilot은 구조화된 답변과 빠른 검색 속도로 긍정적인 평가를 받았다. 그러나 어색한 번역체, 좁은 글씨 간격과 글씨체가 가독성이 떨어진다는 문제점이 있었다.

유용성 측면에서 Naver CUE:는 네이버 쇼핑 링크 혹은 유튜브 링크와 같이 부가적으로 주는 정보가 유용하다는 평가를 받았으며, 정보를 얻는데 시간을 절약해 준다고 느꼈다. 그러나 기존의 정보를 그대로 가져오는 경향이 있어 정보의 깊이가 낮게 느껴졌으며, 제공되는 정보의 양에 대한 불만도 있었다. Bing Copilot은 다양한 정보를 신속하게 얻을 수 있어 리서치를 빠르게 하는데 유용하다는 평가를 받았고, 예상하지 못한 유용한 답변을 제공하기도 했다. 그러나 중복된 답변을 제공하는 경우가 있었고, 원하는 정보를 얻을 때까지 재차 질문해야 하는 불편함이 지적되었다.

표 7. '유용성' 변수에 대한 인터뷰 결과
Table 7. Interview results for variable 'Useful'

Variable	Service	Positive Comments
Useful	Naver CUE:	• It is useful because it provides additional information such as related links. • Saves time obtaining information.
	Bing Copilot	• I can quickly obtain a variety of information. • It is useful because I can get unexpected answers.
	Service	Negative Comments
	Naver CUE:	• It feels like a general answer, so the depth of information is shallow. • The amount of information provided is small.
	Bing Copilot	• There are cases where duplicate answers are provided. • It wasn't the information I wanted, so I had to ask again.

표 8. '신뢰성' 변수에 대한 인터뷰 결과
Table 8. Interview results for variable 'Credible'

Variable	Service	Positive Comments
Credible	Naver CUE:	• The source is indicated as a link to an authorized site, which increases trust. • It provides information that well reflects the conditions I want.
	Bing Copilot	• Provide the source for each piece of information clearly. • It is highly reliable as it provides information based on various sources.
	Service	Negative Comments
	Naver CUE:	• There are times when it gives strange answers. • Blog links feel like advertisements and are therefore unreliable.
	Bing Copilot	• There are cases where incorrect information was provided. • There are times when the links provided do not have the most up-to-date information.

신뢰성에 대해서 Naver CUE:는 사용자가 원하는 조건에 부합하는 답변을 제공하며, 출처를 명확하게 표기하여 신뢰성을 높이는 장점이 있었다. 그러나 맥락에 맞지 않는 답변을 할 때가 있었고, 일부 블로그 링크가 광고성 정보로 느껴지는 경우가 있었다. Bing Copilot은 정보마다 명확한 출처를 제공하고, 출처가 다양하여 신뢰가 간다고 밝혔다. 하지만 일부 연구 참여자는 정확하지 않은 정보를 제공받거나 최신 정보가 아닌 링크를 제공받아 신뢰성이 떨어진다고 지적하기도 했다.

표 9. '매력성' 변수에 대한 인터뷰 결과
Table 9. Interview results for variable 'Desirable'

Variable	Service	Positive Comments
Desirable	Naver CUE:	• It is friendly to Korean, so conversation is natural. • The tone of speech is soft. • The design is clean and simple.
	Bing Copilot	• The use of emoticons makes you feel friendly. • It uses humorous expressions.
	Service	Negative Comments
	Naver CUE:	• It felt like I was talking to an AI pretending to be a human. • The design was simple but not attractive.
	Bing Copilot	• The design is not beautiful. • The translated Korean grammar is awkward. • Pretending to be humorous makes you less attractive.

표 10. '윤리성' 변수에 대한 인터뷰 결과
Table 10. Interview results for variable 'Moral'

Variable	Service	Positive Comments
Moral	Naver CUE:	• Maintain a neutral stance on political issues. • It is fair because it explains both the pros and cons.
	Bing Copilot	• It provides an unbiased answer. • The service does not provide opinions, but leaves the user to make the decision. • It does not feel like it provides advertising information.
	Service	Negative Comments
	Naver CUE:	• It is difficult to tell whether the blog source is an advertisement or not.
	Bing Copilot	• It's inconvenient because they block all political questions. • When recommending information about a specific service, it feels like advertising.

매력성 측면에서 Naver CUE:는 자연스러운 한국어 구사와 깔끔한 디자인으로 긍정적인 평가를 받았지만, 반면에 AI가 사람을 모방하는 것 같아 불쾌하다는 의견이 있었으며, 심플한 디자인이 오히려 매력이 느껴지지 않는다는 평가가 있었다. Bing Copilot은 친근한 말투와 이모티콘 사용으로 사

용자에게 친근감을 주려는 노력이 돋보였으나, 한국어 문법의 어색함과 디자인의 부족함이 단점으로 지적되었다.

윤리성에 측면에서는 두 서비스 모두 중립적인 답변을 제공하도록 설계되었다는 것을 알 수 있었다. 또한, 연구 참여자 대부분이 서비스가 광고성 정보를 노출하지 않고 있다고 느꼈다. 그러나 Naver CUE:는 블로그 링크가 광고성 정보인지 아닌지 불확실하여 광고성 정보가 사용자에게 노출되고 있는 것은 아닌지 의심된다는 의견이 있었고, Bing Copilot은 정치적 질문에 대한 검열로 인해 원하는 답을 얻을 수 없었다는 불편함과 특정 서비스의 링크를 반복해서 제공할 시에 해당 서비스를 광고하는 링크로 인식된다고 밝혔다.

4-5 분석 결과에 대한 논의

리커트 척도를 분석한 결과, 사용성, 유용성, 신뢰성, 윤리성에서는 유의미한 차이가 나타나지 않았으나, 매력성에서는 Naver CUE:가 더 높은 점수를 얻어 유의미한 차이가 있었다. 이러한 결과의 원인을 사용자 경험 인터뷰 결과를 통해 분석하고자 한다.

사용성에서 큰 차이가 없었던 이유는 두 서비스 모두 사용자에게 가장 익숙한 ChatGPT와 사용법이 유사하여 조작이 직관적이었기 때문이다. 유용성 측면에서는 Naver CUE:가 부가적으로 제공하는 정보에 대한 만족도가 높았던 반면, Bing Copilot은 빠른 처리 속도와 참신한 정보 제공에서 강점을 보였다. 두 서비스 모두 유용성 면에서 긍정적인 평가를 받았기 때문에, 만족도 차이가 크지 않았던 것으로 분석된다. 신뢰성에서는 대부분의 사용자가 두 서비스에서 거짓된 정보를 발견하지 못했으며, 출처의 신빙성에 대한 의견 차이가 크지 않았기에 신뢰성 점수에서 유의미한 차이가 없는 것으로 보인다. 그러나 신뢰성에 대한 사용자 경험 인터뷰 결과에서 Naver CUE:가 사용자들이 찾고자 하는 정보의 조건을 더욱 충실하게 반영한 답변을 제공하는 것을 알 수 있었다. Naver CUE:는 25명 중 15명이 원하는 조건을 누락하지 않은 정보를 제공했다고 답변한 반면, Bing Copilot은 25명 중 7명만이 조건을 누락하지 않은 정보를 제공했다는 평가를 내렸다. 윤리성에서도 두 서비스 간 유의미한 차이가 없었는데, 이는 두 서비스 모두 중립적인 답변을 제공하며, 대부분의 연구 참여자들이 광고성 정보를 제공하지 않는다고 느꼈기 때문이다. 이를 통해 두 서비스가 모두 인공지능 윤리 정책에 충실하게 설계되어 윤리적으로 문제를 일으킬 가능성을 최소화하고 있음을 확인할 수 있었다.

매력성 측면에서는 리커트 척도 평가에서 유의미한 차이가 나타났으며, Naver CUE:가 Bing Copilot보다 더 높은 점수를 얻었다. 이러한 결과는 주로 UI 디자인의 매력성에서 큰 차이를 보였기 때문이다. “서비스의 디자인이 심미적인가?”라는 질문에 대해 Naver CUE:는 25명의 응답자 중 23명이 "디자인이 깔끔하고 정돈되었다"고 긍정적으로 평가하였다. 반면, Bing Copilot은 같은 질문에 대해 25명 중 23명이 부정적인 답변을 하였으며, 이들은 Bing Copilot의 UI 디자인과 폰트에서 불편함을 느끼고, 한국인 사용자에게 적합하지 않다고 판단하였다. 이러한 명확한 반응 차이로 인해 Naver CUE:의 매력성 점수가 더 높게 평가된 것으로 분석된다.

사용자 경험 인터뷰를 통해 두 서비스의 장점과 단점을 분석하고, 이에 따른 개선 방안을 도출할 수 있었다. Naver CUE:의 주요 장점은 한국인 사용자에게 최적화된 서비스라는 인식이다. 이는 네이버가 한국에서 가장 널리 사용되는 검색 엔진이라는 점에서 기인하며, UI 디자인과 폰트가 한국인에게 익숙한 형태로 제공된다. 또한, 네이버 블로그나 네이버 쇼핑 링크와 같은 출처를 기반으로 한 정보 제공은 사용자에게 친숙하게 다가와, 한국인에게 최적화된 서비스라는 평가를 받았다. 반면, Naver CUE:의 개선점으로는 사용자에게 부적절한 답변이 제공될 수 있다는 점이 지적되었다. 예를 들어, 멕시코 현지 맛집을 검색했을 때, 국내의 멕시코 음식점을 추천하거나, 20대 남성 향수를 추천받았지만 20대 여성용 향수가 포함된 경우가 있었다. 이러한 문제를 예방하기 위해, 사용자가 질문한 핵심 맥락과 키워드를 보다 정확히 판단하고, 답변 제공 시 핵심 키워드에 부합하지 않는 정보는 필터링하는 시스템의 강화가 필요하다.

Bing Copilot의 주요 장점은 GPT-4 기반의 빠른 성능과 뛰어난 정보 정리 능력이다. 그러나 개선이 필요한 점으로는 사용자 요구 조건을 보다 정확하게 반영하는 능력 향상과 UI 디자인 개선이 있다. 정보 검색을 목적으로 사용할 때, 사용자가 원하는 조건이 반영되지 않아 필요한 정보를 얻지 못할 경우, 서비스의 만족도가 크게 저하될 수 있다. 이에 대한 개선 방안으로 Naver CUE:의 정보 도출 과정을 참고할 수 있다. Naver CUE:는 신뢰성이 부족하거나 정확하지 않은 답변을 줄이기 위해서 RAG(Retrieval-Augmented Generation) 방식을 기반으로 응답을 생성한다. RAG는 미리 학습된 LLM과 외부 데이터를 결합해 응답을 생성하는 방식으로, 외부 데이터를 검색해 문맥을 활용함으로써 환각 현상을 방지한다[18]. Naver CUE:는 Reasoning 기술을 통해 질문의 본질을 분석하고 해석하며, Evidence Selector 기술로 다양한 검색 결과 중 가장 관련성 높은 정보를 선별하고, Factually Consistent Generation 기술을 통해 정보의 출처와 신뢰성을 검증하는 단계를 거쳐 응답을 생성한다[19]. 이를 참고하여 Bing Copilot은 GPT-4 기반에 더하여, 검색 결과의 정확성과 품질을 높이기 위해 검색 기능을 위한 추가적인 모델을 구축하고, 단계별 검색 결과 향상 프로세스를 도입할 것을 제안한다. 또한, 다국적 사용자들의 만족도를 높이기 위해 범용성과 심미성을 고려한 자체 서체를 개발해 UI 디자인에 적용할 것을 제안한다. 한 연구에 따르면, 앱 디자인의 심미성은 유용성에 대한 인식을 높이고, 사용 용이성과 인지된 즐거움을 증대시킴으로써 서비스에 대한 충성도를 높이는 데 중요한 역할을 한다는 점에서, 심미성의 고려가 필수적이다[20].

V. 결 론

본 연구는 정보 검색 분야에서 AI를 활용한 서비스인 Naver CUE:와 Bing Copilot이 사용자가 정보를 검색하는 경험에서 AI가 효과적인지 파악하기 위한 연구를 진행했다. 평가 요소는 사용성, 유용성, 신뢰성, 매력성, 윤리성이었으며, 각 평가 요소마다 리커트척도 평가와 사용자 경험 인터뷰를 진행했다.

리커트 척도 평가 결과를 Wilcoxon 부호순위 검정으로 분석한 결과, 사용성, 유용성, 신뢰성, 윤리성에서는 유의미한 차이가 나타나지 않았다. 그러나 매력성에서는 유의미한 차이가 있었으며, 그 원인은 사용자 경험 인터뷰를 통해 파악할 수 있었다. Naver CUE:는 깔끔한 디자인 덕분에 매력성 면에서 Bing Copilot보다 더 높은 점수를 얻었다. 이를 통해 UI 디자인의 심미성이 서비스의 매력성에 영향을 미칠 수 있음을 확인할 수 있다.

본 연구는 표본 크기 문제와 실험 환경의 제약에 대한 한계가 존재한다. 표본이 20-30대의 한국인으로만 구성하였으며, 샘플 사이즈가 작다는 한계점을 가지고 있다. 이러한 제한으로 인해 다른 연령대 및 국적의 서비스 사용자에게는 다른 결과가 나타날 수 있다. 또한, 실험 연구를 진행하여 샘플 크기가 작아 통계 검정의 검정력이 낮아져, 정규성을 평가하는 데 더 많은 변동성이 발생할 수 있고, 제1종 오류(Type I Error) 또는 제2종 오류(Type II Error)의 위험이 높아 오류가 존재할 수 있다. 실험 환경의 제약으로는 모바일 환경에서의 연구가 진행되지 않아 실제적 유용성에 제한이 있어 모바일 환경에 대한 후속 연구가 필요하다.

그럼에도 불구하고 본 연구는 AI 검색 서비스에 특화된 평가 요소를 구축하였고, 실제 사용자 경험을 기반으로 서비스의 장단점과 개선 방안을 도출했다는 점에서 중요한 의미를 가진다. LLM 기반의 AI 서비스가 정보 검색 분야에서 발전하고 있는 현시점에서, 본 연구가 서비스의 장점을 극대화하고 단점을 극복하는 데 기여하여 AI 검색 서비스의 대중화에 도움이 되기를 기대한다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발 사업으로 수행되었음(과제번호 : RS-2024-00441262, 기여율: 100%).

참고문헌

[1] M. S. Choi, T. D. Kim, and S. E. Meong, 2023 Survey on the Internet Usage, National Information Society Agency, Daegu, NIA VIII-RSE-C-23069, April 2024.

[2] AI Times. Search Paradigm Shift [Internet]. Available: https://www.aitimes.com/news/articleView.html?idxno=150570.

[3] Google Korea Blog. [I/O 2024] Google's Generative AI Search: Leave the Search to Google [Internet]. Available: https://blog.google/intl/ko-kr/products/explore-get-answers/generative-ai-google-search-may-2024-kr/.

[4] NewsQuest. "From Simple Search to Questions"...'The Eve of the Storm' in the Search Engine Market, Creating a 'New Version' with Generative A [Internet]. Available: https://www.newsquest.co.kr/news/articleView.html?idxno=229238.

[5] Naver Channel Tech. Naver Generative AI Search, Cue [Internet]. Available: https://channeltech.naver.com/contentDetail/43.

[6] H. Park, "Hallucination Issues and Ethical Challenges of Generative AI: Focusing on Topics Applicable to Elementary AI Ethics Education," *Korean Journal of Elementary Education*, Vol. 34, No. 4, pp. 21-36, December 2023. https://doi.org/10.20972/kjee.34.4.202312.21

[7] J. Kwon, "Conditions for Platform-Based Innovative Technologies to Become Universally Popular: The Chasm Must Be Overcome," *Korean Association of Converging Business Review*, No. 59, pp. 10-19, May 2024.

[8] J. Y. Kim and H. J. Choi, "A Study on Robot UX Design for Human-Robot Interaction," *Design Research*, Vol. 9, No. 3, pp. 475-485, September 2024. https://doi.org/10.46248/kidrs.2024.3.475

[9] J. Nielsen, *Usability Engineering*, San Diego, CA: Academic Press, 1993. https://doi.org/10.1016/C2009-0-21512-1

[10] S. H. Lee and S. I. Kim, "A Comparative Study on LLM-Based AI Service User Experience -Focusing on ChatGPT and Clova X-," *Journal of Cultural Product & Design*, Vol. 75, pp. 11-22, December 2023.

[11] V. Vasantham, K. Preetham, G. Pavan Kumar, L. Krishna, and K. Sandeep, "Combination of Scrum Lean-UX-based AI UX Design," in *Proceedings of the 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, pp. 1372-1378, July 2023. https://doi.org/10.1109/ICESC57686.2023.10193361

[12] S.-J. Hur, J. Youn, and S.-H. Kim, "A Proposed Framework for UX Evaluation of Artificial Intelligence Services," in *Proceedings of KIICE General Spring Conference*, Yeosu, pp. 274-276, May 2021.

[13] J. Y. Jeong and K. K. Chang, "A Study on the Usability

Evaluation of Laundry O2O Service Reflecting the Living Pattern of One-Person Households -Focused on Laundrygo and Getwashswat Application-," *Journal of Basic Design & Art*, Vol. 22, No. 5, pp. 469-480, October 2021. https://doi.org/10.47294/KSBDA.22.5.35

[14] S. H. Hwang and D. Y. Ju, "Usability Evaluation of Artificial Intelligence Search Services Using the Naver App," *Science of Emotion & Sensibility*, Vol. 22, No. 2, pp. 49-58, June 2019. https://doi.org/10.14695/KJSOS.2018.22.2.49

[15] M. J. An and T. Y. Kang, "A Study on the Evaluation of ChatGPT User Experience Design for Digital Transformation Management -Cross-Utilization of OpenAI ChatGPT and Microsoft Bing ChatGPT-," *Journal of the Korean Society Design Culture*, Vol. 29, No. 2, pp. 237-247, June 2023. https://doi.org/10.18208/ksdc.2023.29.2.237

[16] Google Korea Blog. 100 Key Announcements at Google I/O 2024 [Internet]. Available: https://blog.google/intl/ko-kr/products/google-io-2024-100-announcements-kr/.

[17] S. Park, "Evaluation of Mobile-Based Web Search Services: Suggestions for Needed Improvements," *Journal of the Korean Society for Library and Information Science*, Vol. 49, No. 4, pp. 317-334, November 2015. https://dx.doi.org/10.4275/KSLIS.2015.49.4.317

[18] J. Kim, Research on Data Chunking Strategies for Enhancing LLM Service Quality Using the RAG Technique, Master's Thesis, Korea University, Seoul, February 2024. https://www.doi.org/10.23186/korea.00000279092.11009.0000389

[19] H. Y. Yu, S. H. Yu, and Y. B. Kim, "NAVER Cue: Search Service Based on Large Language Models," *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 41, No. 11, pp. 34-41, November 2023.

[20] M. J. Kim, S. K. Yoon, and J. H. Choi, "The Effect of Aesthetics on the User's Loyalty of Mobile Video Streaming Apps," *Journal of the Korean Society of Design Culture*, Vol. 20, No. 4, pp. 63-74, December 2014.

박창준(Changjun Park)

2016년 : 숭실대학교
글로벌미디어학부 (공학사)

2016년～현 재: 숭실대학교 글로벌미디어학부 학사과정 수료
※관심분야 : 서비스 기획, UXUI, 생성형 AI

이정진(Jungjin Lee)

2010년 : 숭실대학교 미디어학부 학사
2012년 : KAIST 문화기술대학원 석사
2017년 : KAIST 문화기술대학원 박사

2016년～2020년: (주)카이 연구이사
2020년～현 재: (주)라이브커넥트 CTO 사외이사
2020년～현 재: 숭실대학교 글로벌미디어학부 조교수
※관심분야 : 컴퓨터 그래픽스, VR/AR, 몰입형 시각 미디어, 이미지/비디오 응용, HCI 등