Proposal of User Interface Based on Heavy User Usage Analysis in LLM Service

Jong Hyun Chin[1], So Young Lee[1], Chae Lin Park[1], Myeong Heum Yeoun[2*]

[1]Department of Smart Experience Design, Student, Kookmin University, Seoul, Korea

[2]Department of Smart Experience Design, Professor, Kookmin University, Seoul, Korea

Abstract

Background The use of Large Language Models (LLMs) is growing in various industries, including reasoning, question answering, and code generation. However, the overall direction and analysis of LLM research is focused on the availability and efficiency of service systems, and interface research based on behavioral analysis of users in services is lacking.

Methods This study aims to define the different usage behaviors and underlying mental models of users in LLM services by proposing a quantitative positioning map through pre-survey data and defining users into three groups, and then proposing and verifying improvement strategies through service usage screen analysis and in-depth interviews.

Results Among the three groups defined through the survey, we defined a mental model that the target group, Heavy users, has a higher expectation level of conversation quality through LLM service compared to the comparison groups, Middle and Novice users, through the main behaviors of prompt usage and opening a new conversation window, and proposed three improvement strategies (1. Proposing a hierarchy of history and selecting a range of conversation information, 2. Storing and processing conversations with contextual information, and 3. Providing prompt guidelines and proposing a function to modify answers).

Conclusions For Improvement Strategy 2, the perceived usefulness (PU), a measure of the Technology Acceptance Model (TAM), was significant, but the perceived ease of use (PEU) was not, which was confirmed in the post-interview due to the limited sample of target users during the test and the lack of activeness to drive the enhancement Design. On the other hand, improvement strategies 1 and 3 showed statistical significance for both PU and PEU, and positive feedback was confirmed in the post-interview.

Keywords AI, Large Language Model, User Interface

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A03084950)

*Corresponding author: Myeong Heum Yeoun (yeounmh@kookmin.ac.kr)

Citation: Chin, J., Lee, S., Park, C. L., & Yeoun, M. H. (2024). Proposal of User Interface Based on Heavy User Usage Analysis in LLM Service. *Archives of Design Research, 37*(4), 287-313.

http://dx.doi.org/10.15187/adr.2024.08.37.4.287

Received : Apr. 30. 2024 ; **Reviewed :** Jun. 25. 2024 ; **Accepted :** Jun. 28. 2024
pISSN 1226-8046 **eISSN** 2288-2987

1. 서론

1. 1. 연구배경 및 목적

지난 몇 년간 거대 언어 모델(Large Language Model)은 Chat GPT, CLOVA X와 같은 LLM 서비스의 형태로 인간이 콘텐츠에 접근하거나 제작하는 방식에 큰 변화를 만들어 내고 있다(Alexandre, 2023). 특히 논리적 추론, 질문 응답, 코드 생성 등 다양한 분야에서 효과적인 방향을 일으키며 빠르게 성장 중에 있다.(Y chen et al,.,2023). 밸류에이츠 리포트(Valuates Reports)가 2023년 6월 발표한 '전 세계 대규모 언어 모델(LLM) 시장 조사 보고서'에 따르면, 2022년 105억 달러였던 LLM 시장 규모가 연평균 성장률 21.4%로 성장해 7년 후 408억 달러에 이를 것이라 전망하였다.(Yi, Seohoo, 2023). 이와 같이 거대 언어 모델인 LLM은 시장 전반에서의 활용이 확대되고 있으며, 특히 다양한 업무 맥락에서 그 효용성을 입증한 바 있다. Tyna(2023)의 연구 결과에 따르면, LLM 서비스는 약 80% 미국 근로자의 작업에 최소 10% 이상 영향을 미칠 것이라고 예상하며, 그 중 약 19%의 근로자는 작업의 최소 50% 이상 영향을 받을 것이라고 전망한다. 때문에 업무 맥락을 중심으로 사용자들의 LLM 서비스 사용 행태에 대해 심도 깊이 분석하는 연구는 매우 가치 있을 것이라 판단된다.

Figure 1 LLM Positioning Map of LLM Studies

선행 연구에 대한 분석을 통한 LLM 연구의 기회요인을 정의하기 위하여, 2023년 8월부터 2024년 1월까지 발행된 38건의 국내외 HCI 및 디자인 연구를 포지셔닝 맵을 통해 정리하였다. X축은 연구의 결과 및 목적(신규 경험을 제안하는가, 기존의 경험을 분석 및 개선하는가)을 기준으로, Y축은 연구 대상(LLM 내부 시스템을 분석하는가, 사용자의 서비스 접점 내 경험을 분석하는가)으로 정의하여 Figure 1과 같이 표현하였고, LLM 분야의 연구가 크게 3가지 방향성으로 이루어지고 있는 것을 확인하였다.
2사분면에 위치한 'Discovering Factor'는 LLM 서비스 접점 내 특정 사용 경험의 주요 원인 및 요인을 정의하는 방향의 연구로, IT계열 종사자들의 LLM 서비스 사용 인식을 TAM을 통해 정의한 연구(Agossah, 2023), 또는 전문가가 아닌 일반 사용자가 프롬프트 작성에 어려움을 겪는 요인을 디자인 프로브를 통해 밝혀낸 연구(Wong, 2023) 등이 대표적이다.
3사분면에 위치한 'System Develop'은 LLM 시스템 구조에 대한 분석 및 실험을 통해 개선하는 연구로, 토큰(LLM 대화의 단위) 차원에 따라 데이터 레이어를 수정 및 실행하여, 단어 추론 능력을 극대화한 연구(Wang, J, & Yang X, 2023), 또는 LIBRO 프레임워크 제안을 통해 버그 재현 테스트 성능을 개선한 연구(Yoon, J, 2023)가 대표적이다.
4사분면의 'Prompt Discovering'은 LLM을 시스템을 통해 신규 서비스 및 기능의 가용성을 확인 후 제안하는 연구로, LLM 서비스를 통해 대중 연설 연습용 시뮬레이터 및 연설자를 생성한 연구(Park, J & Choi, D, 2023)와 디자인 아이디에이션의 확장 및 결합을 언어 모델을 통해 구현한 연구(Giulia et al., 2022)가 대표적이다.

전반적인 연구의 대상이 사용자 경험보다는 서비스 시스템의 가용성 및 효율성에 집중되어 있음이 확인되었으며, 특히 LLM 서비스 접점 내 사용자 조사를 통해 신규 경험을 제안하는 연구(1사분면)는 미비하다 판단된다. 하지만 성공적인 기술 경험의 조건으로 사용자의 맥락적 요구와의 조화를 주장한 Shneiderman. B.(2002)와 같이 서비스 내 사용자 행태 기저의 니즈에 대한 분석을 기반한 기술 경험을 제안하는 1사분면의 연구는 매우 중요할 것이라 판단된다. 때문에 본 연구는 앞서 언급한 업무 맥락 중심의 사용자 경험 연구에 목적을 두어, LLM 서비스의 사용 능력과 기술 수용 주기 모델(TAM)을 통해 세분화된 3가지 사용자 집단 간의 사용 행태를 분석하고, 분석 과정에서 획득한 인사이트를 바탕으로 경험 개선을 위한 신규 인터페이스와 기능을 제안한 뒤 실효성을 검증하고자 한다.

1. 2. 연구 프로세스

문헌 연구를 통해 'TAM(Technology Acceptance Model)'과 '서비스 사용 능력'을 LLM 서비스 사용자를 세분화하기 위한 두 가지 기준으로 수립하였다. 이를 바탕으로 7점 Likert 척도 기반의 사전 설문을 수행하여 인터뷰이들을 포지셔닝 맵 내에 정량적으로 배치하였다. 이후 사전 설문의 개괄적인 이용 행태 문항 중 '업무 및 학습 맥락 내 LLM 서비스 의존도'를 기준으로 인터뷰이들을 3가지 그룹으로 정의하였다.
세분화된 세 그룹의 이용 행태 및 목적을 분석하기 위해 사용자들의 서비스 이용 화면을 이미지로 취합하여 대화 중 프롬프트의 활용 및 대화 히스토리 개설 행태를 분석하였고 해당 분석을 통해 정의된 집단별 주요 행태 기저의 멘탈 모델을 이해하기 위해 총 20명을 대상으로 In-depth 인터뷰를 진행하였다.
인터뷰 인사이트를 기반으로 도출된 UX 전략을 인터페이스 레벨로 구체화한 뒤 영상으로 제작하였고 이후 정규성을 만족하는 33명의 테스트 대상자를 표집하여 전후 경험의 개선 정도를 정량적, 정성적으로 분석하기 위한 사용성 테스트를 진행하였다. 최종적으로 개선안의 유의성을 검증하기 위하여 통계검정을 수행하였다.

Figure 2 Research Model

2. 사용자 세분화 기준 수립을 위한 문헌 조사

2. 1. TAM (기술 수용 모델)

TAM(Technology Acceptance Model)은 새로운 기술이 개인의 사용에 영향을 미치는 요인을 탐구하기 위한 모델로 합리적 행위 이론(Theory of Reasoned Action)을 기반으로 구축된 모델이다. 텍스트 프로세서, 스프레드시트 애플리케이션, 이메일, 웹 브라우저, 웹사이트 등과 같은 언어 기반의 신규 기술에 대한 사용자의 수용 의도를 설명하는 데 유용한 척도로 언급되어 오고 있다(JR Lewis, 2017).

이 모델은 특정 시스템을 사용하는 것이 어렵지 않다고 개인이 믿는 정도를 의미하는 인지된 사용용이성(Perceived Ease of Use, PEU)과 특정 시스템을 이용하는 것이 자신의 업무 성과를 개선시킬 것이라고 개인이 믿는 정도를 의미하는 인지된 유용성(Perceived Usefulness, PU), 이 두 가지 변수로 구성되어 있다(FD Davis, 1989; Kim, 2017).

Davis(1989)는 두 신념 변수가 사용자의 정보 기술 이용 태도 및 행동 의도와 큰 관련성을 가짐을 발표하며 특정 서비스에 대한 사용자의 적극성을 설명하는 척도가 될 수 있음을 시사한다. 즉 해당 모델을 통해 측정된 인지된 유용성과 사용 용이성이 높을수록 서비스 내에서 지속적으로 제시되는 신규 기능과 경험에 대해 적극적인 사용자일 것이라는 의미로 해석될 수 있고 이는 서비스 내 사용자의 행태를 정의할 수 있는 중요한 기준이 될 것이다. 따라서 본 연구에서는 LLM 서비스 내 사용자의 적극적 이용 태도에 대한 기준인 TAM을 사용자 세분화의 하나의 축으로 정의하였다.

2. 2. 서비스 사용능력

Hansen(1971)이 발표한 User engineering principles for interactive systems의 첫 번째 원칙이 'know the user'인 만큼 서비스 시스템을 제작할 때 사용자들이 여러 맥락을 고려하여 작동하기 쉽도록 설계하는 것에 목적을 두곤 한다. 하지만 사용자들의 배경과 지식은 문화적 차이뿐만 아니라 개인적 차이에서도 크게 나타나기 때문에 동일하게 설계된 시스템 안에서도 배경지식, 인지 능력, 시스템 사용과 관련된 지식 등을 나타내는 서비스 사용 능력에 따라 상이한 사용 행태를 보일 것이라고 생각할 수 있다(Jakob, 2018).

특히 답변 및 산출된 결과를 통제하기 어려운 LLM의 경우 서비스 사용 능력에 따른 경험의 격차는 더욱 심화될 것이라 판단되는데 선행 연구에 따르면 학습한 데이터의 범위를 이해하며(Y, chen, 2023)(Ting ceng, 2023), 대화를 통해 제공하는 정보의 구조인 'prompt'에 대한 이해도에 따라(Pereira, 2023; Breanna et al., 2024; Cho, 2024) 상이한 품질의 대화 경험을 보인다 주장하였다.

이에 따라 LLM 서비스의 사용 능력은 'LLM이 학습한 데이터의 범위와 종류를 이해하여 구조화된 Prompt 기반의 대화를 수행할 수 있는 능력'이라 판단된다. 본 연구에서는 위와 같은 사용 능력에 따라 세분화된 사용자의 서비스 경험을 분석하고자 또 하나의 축으로 정의하였다.

3. 정성 조사를 통한 사용자 분석

3. 1. 사전 설문 제작

앞서 언급된 2가지 세분화 기준을 바탕으로 7점 Likert 척도의 등간 데이터를 사전 설문을 통해 취합하여 21명의 인터뷰이를 정량적으로 Positioning Map에 배치하여 그룹화하고, 해당 인터뷰이들의 사용 행태 및 목적을 이해하기 위하여 인터뷰 진행 전 사전 설문을 배포하였다.

2차원 Positioning Map의 X축에 해당하는 LLM 서비스 사용 능력은 7개 문항으로 49점 만점이었고, Y축에 해당하는 TAM 문항의 경우 선행 연구에서 제작된 문항을 바탕으로 지각된 유용성(PU) 6문항, 지각된 사용 용이성(PEU) 6문항으로 총 12개 문항, 84점 만점으로 구성하였다. 또한 인터뷰이들의 개괄적인 사용 행태(빈도, 의존도, 만족도)에 대한 문항과 함께, 사용자들의 주사용 목적을 묻는 객관식 문항과 자세한 사용 맥락을 묻는 주관식 문항도 사전 설문에 포함하였다.

사전 설문은 인터뷰 진행 10일 전인 2024. 01. 17.부터 3일간 인터뷰이들에게 배포한 뒤 분석하였다.

Table 1 Pre-Interview Survey

구분		번호	문항	척도	선행 연구
1. 개괄적인 사용 행태 문항		1-1	업무나 학습 환경에서 Chat GPT나 Clover X와 같은 LLM 서비스를 사용함에 있어서, 어느 정도 만족하나요?	7점 LIKERT 척도	
		1-2	업무나 학습 환경에서 Chat GPT나 Clover X와 같은 LLM 서비스를 사용함에 있어서 불만족스러운 경험이 있었다면 상세히 작성 부탁드립니다.	주관식	
		1-3	다음 중 사용하는 LLM(Large Language Model) 서비스를 선택해 주세요_다중 선택 가능	명목 척도	
		1-4	하루에 업무 및 학습 상황에서 Chat GPT와 Clover X와 같은 LLM 서비스를 통해 몇 개 정도의 주제(토픽)로 대화하나요?		
		1-5	하나의 주제(토픽)의 대화에서 평균 몇 번 정도 '턴'의 대화를 진행하나요?		
		1-6	텍스트 기반의 업무 및 학습 상황(읽기, 쓰기 요약하기, 정리하기 등)에서 Chat GPT, Clover X와 같은 LLM 서비스에 어느 정도 의존하나요?	7점 LIKERT 척도	
TAM 문항	2. 지각된 유용성 (PU)	2-1	LLM 서비스를 직장에서 사용하면 작업을 더 빨리 완수할 수 있을 것이라 판단된다.	7점 LIKERT 척도	FD Davis, 1989, JR Lewis, 2017
		2-2	LLM 서비스를 사용하면 직무 성능이 향상될 것이라 판단된다.		
		2-3	직장에서 LLM 서비스를 사용하면 생산성이 증가할 것이라 판단된다.		
		2-4	LLM 서비스를 사용하면 직무의 효과성이 향상될 것이라 판단된다.		
		2-5	LLM 서비스를 사용하면 일을 더 쉽게 할 수 있을 것이라 판단된다.		
		2-6	직장에서 LLM 서비스가 유용할 것이라 판단된다.		
	3. 지각된 사용 용이성 (PEU)	3-1	LLM 서비스를 조작하는 법을 배우는 것이 쉬울 것이라 판단된다.		
		3-2	LLM 서비스를 원하는 대로 작동시키기 쉬울 것이라 판단된다.		
		3-3	LLM 서비스와의 상호 작용이 명확하고 이해하기 쉬울 것이라 판단된다.		
		3-4	LLM 서비스 사용법이 명확하고 이해하기 쉬울 것이라 판단된다.		
		3-5	LLM 서비스 사용에 능숙해지는 것이 쉬울 것이라 판단된다.		
		3-6	LLM 서비스를 사용하기 쉬울 것이라 판단된다.		
4. 서비스 사용 능력 문항		4-1	원활하고 정확한 대화 품질을 위하여 LLM 서비스의 사용 방법을 학습하였는가?		Jakob, 2018,
		4-2	대화 주제와 목적에 따라 좌측 대화창(히스토리)을 관리 및 개설하며 사용하는가?		
		4-3	조금 더 구체적이고 높은 품질의 대화를 위해 플러그인을 활용하는가?		
		4-4	대화 시에 프롬프트(prompt)의 구조를 이해하여, 적은 횟수의 대화턴으로도 만족스러운 결과를 얻는 편인가?		Pereira2023, Breanna et a.l, 2024 Cho,2024
		4-5	LLM 서비스가 학습한 데이터의 한계를 이해하고 대화하는가?		Y, chen, 2023 Ting ceng, 2023
		4-6	답변 생성 시의 속도와 정확도 면에서 유리한 언어를 이해하고 사용하는가?		
		4-7	LLM 서비스가 제작 가능한 답변의 형식의 다양성을 바탕으로(이미지 생성, 도표 제작 등의 대답) 대화하는 편인가?		
5. LLM 서비스 주사용 목적	1순위	5-1	업무나 학습 환경에서 Chat GPT나 Clover X와 같은 LLM 서비스를 가장 많이 사용하는 목적은 무엇인가요?	명목 척도	
		5-2	선택하신 주요 사용 목적에 대한 사용 맥락을 자세히 작성해주세요.	장문형	
	2순위	5-3	업무나 학습 환경에서 Chat GPT나 Clover X와 같은 LLM 서비스를 두 번째로 많이 사용하는 목적은 무엇인가요?	명목 척도	
		5-4	선택한 두 번째 주요 사용 목적에 대한 사용 맥락을 자세히 작성해주세요.	장문형	
	3순위	5-5	업무나 학습 환경에서 Chat GPT나 Clover X와 같은 LLM 서비스를 세 번째로 많이 사용하는 목적은 무엇인가요?	명목 척도	
		5-6	선택하신 세 번째 주요 사용 목적에 대한 사용 맥락을 자세히 작성해주세요.	장문형	

3. 2. 사전 설문 분석

3. 2. 1. 정량적 포지셔닝 맵

사전 설문을 통해 취합한 7점 척도 데이터(서비스 사용 능력과 TAM)를 바탕으로 두 기준 간의 양립 가능성을 확인하기 위하여 상관관계 분석을 진행하였고 상관 계수 값이 0.1779로 매우 낮게 나타나 양립됨을 확인하였다.

측정한 각 기준의 만점을 기준으로 X축 가로(서비스 사용 능력 7문항 X 7점) 49픽셀, Y축 세로(TAM 12문항 X 7점) 84픽셀의 필드를 제작하였고 각 축 데이터의 2Q(2nd Quartile = Median)를 측정하여 포지셔닝 맵 필드 내의 축의 위치(X축 2Q: 26 / Y축 2Q: 62)를 정의하였다. 이후 인터뷰이의 서비스 사용 능력과 TAM 측정값을 바탕으로 포지셔닝 맵에 배치하였다.

최종적으로 배치된 인터뷰이들의 개괄적인 사용 행태 문항[Table 1. 1-1 ~ 1-6] 및 주 사용 목적 문항[Table 1. 5-1 ~ 5-5]을 기준으로 사분면별 또는 포지셔닝 맵 내에서의 거리를 기준으로 분석하고자 하였다.

3. 2. 2. 포지셔닝 맵과 사전 설문을 통한 Group 정의

Table 2 Cross Analysis between Positioning Map and Service Use Behaviors Questionnaire

그룹	인터뷰이	의존도[1-6]	만족도[1-1]	1일 대화 수	대화당 턴수
Novice User Group	i-12	**(1점) 14.29%**	(3점) 42.86%	약 3개 미만	약 10회 미만
	i-14	**(1점) 14.29%**	(2점) 28.57%	1개 미만	약 5회 미만
	i-4	**(2점) 28.57%**	(6점) 85.71%	약 10개 미만	약 5회 미만
	i-9	**(2점) 28.57%**	(5점) 71.43%	약 3개 미만	약 15회 미만
	i-13	**(2점) 28.57%**	(4점) 57.14%	1개 미만	약 15회 미만
	i-11	**(2점) 28.57%**	(4점) 57.14%	약 3개 미만	약 10회 미만
	i-16	**(2점) 28.57%**	(2점) 28.57%	약 3개 미만	약 5회 미만
	i-19	**(2점) 28.57%**	(6점) 85.71%	약 3개 미만	약 10회 미만
Middle User Group	i-3	**(3점) 42.86%**	(5점) 71.43%	약 3개 미만	약 10회 미만
	i-15	**(3점) 42.86%**	(6점) 85.71%	약 3개 미만	약 10회 미만
	i-17	**(4점) 57.14%**	(5점) 71.43%	약 5개 미만	약 20회 미만
	i-18	**(3점) 42.86%**	(5점) 71.43%	1개 미만	약 15회 미만
Heavy User Group	i-1	**(5점) 71.43%**	(6점) 85.71%	약 3개 미만	약 20회 미만
	i-2	**(5점) 71.43%**	(6점) 85.71%	약 5개 미만	약 5회 미만
	i-6	**(5점) 71.43%**	(6점) 85.71%	약 5개 미만	약 15회 미만
	i-5	**(5점) 71.43%**	(3점) 42.86%	1개 미만	약 10회 미만
	i-21	**(6점) 85.71%**	(7점) 100.00%	약 10개 이상	약 10회 미만
	i-22	**(6점) 85.71%**	(7점) 100.00%	약 10개 이상	약 5회 미만
	i-8	**(7점) 100.00%**	(7점) 100.00%	약 10개 이상	약 10회 미만
	i-10	**(7점) 100.00%**	(7점) 100.00%	약 10개 이상	약 5회 미만

포지셔닝 맵 내 인터뷰이의 위치를 기반하여 사용자 세분화를 위한 기준을 정의하기 위해 Table 1의 개괄적 사용 행태 및 목적 문항[만족도(1-1), 일일 대화 개수(1-4), 대화당 턴수(1-5), 서비스 의존도(1-6), 서비스 주 사용 목적(5-1 ~ 5-5)]의 데이터와 포지셔닝 맵 내 인터뷰이들의 위치를 교차 분석하였다[Table 2].

분석 결과 Figure 3의 결과와 같이 포지셔닝 맵에서 우상향에 위치할수록 '업무 및 학습 맥락 내 LLM 서비스 의존도' 문항[Table 1_문항 번호 1-6]의 측정값이 높아지는 경향성을 확인하였다. 즉 TAM을 통해 측정된 서비스 사용의 적극도와 서비스 사용 능력이 높은 사용자일수록 LLM 서비스 의존도 또한 높다는 결과를 확인할 수 있었다.

이를 통해 의존도 점수[Table 1_문항 번호 1-6]를 기반한 포지셔닝 맵 내 2개의 우하향 대각선을 배치하였고 오른쪽 영역의 집단을 Heavy user, 중간에 위치한 인터뷰 집단을 Middle user, 좌측에 위치한 인터뷰이 집단을 Novice user 집단으로 정의하였다.

본 연구에서는 세 그룹 중 실제 서비스에서 결제 및 리텐션 등의 활동성 지표에 가장 큰 영향을 미치는 서비스 내 주요 집단이라 판단되는 Heavy 유저를 메인 타깃으로 선정하였고 나머지 두 비교군 집단 간의 분석을 통한

인사이트를 얻고자 하였다. 이후 그룹을 세분화한 기준 척도인 '업무 및 학습 맥락 내 LLM 서비스 의존도'가 서비스 사용 행태에 어떠한 차이를 보일 것인지에 대한 추가 분석을 진행하고자 하였다.

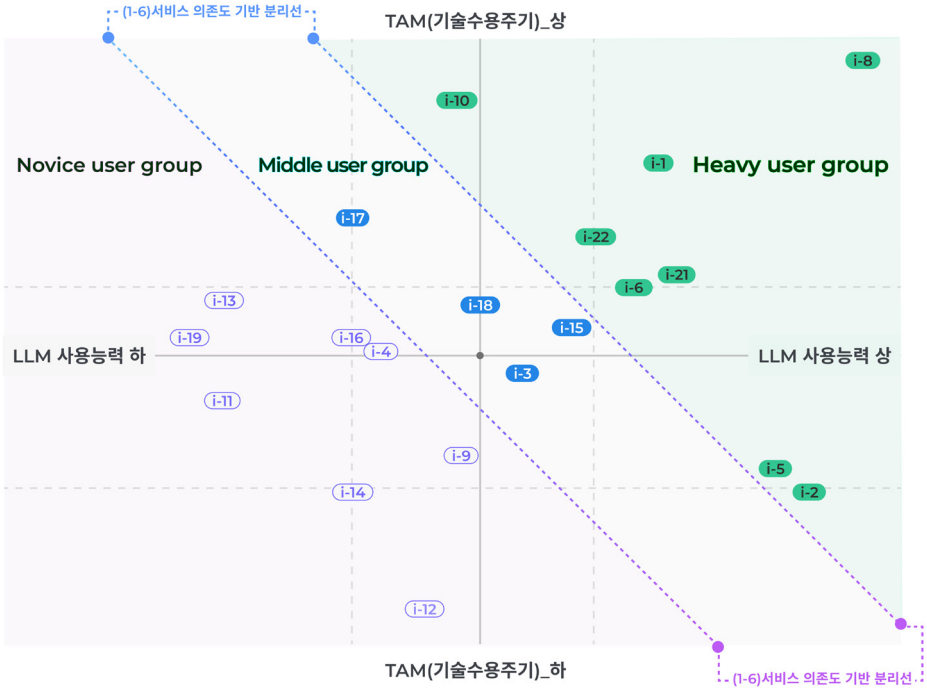

Figure 3 Cross Quantitative Positioning Map for User Segmentation

3. 3. LLM 사용 화면 분석

집단을 세분화한 기준인 서비스 능력 및 TAM을 바탕으로 업무 및 학습 맥락 내 LLM 서비스 사용 의존도 차이에 따라 서비스 사용 행태가 어떠한 차이를 보일지에 대한 분석과 더불어, 사용자들의 대화 목적을 유형화하기 위하여 2024.01.22.부터 2일간 21명의 인터뷰이들에게 사용 화면에 대한 캡처 이미지를 요청하였고, Heavy user 7명, Middle user 5명, Novice user 6명, 총 18명으로부터 이미지 데이터를 취합할 수 있었다. 한 유저당 최소 3~4개 이상의 주 사용 목적 대화에 대한 이미지를 취합하는 방식으로, 약 146장의 이미지가 취합되었다.

[Figure1] 카카오톡을 통한 첨부 방식 예시 이미지

Figure 4 LLM Service Usage Screen Shoot Request

3. 3. 1. LLM 서비스 사용 목적의 유형화

Table 3 Conversation Types in the LLM Service

구분		T 1. 자신을 위한 정보 탐색 및 정리	T 2. 타인, 양식화를 위한 정보 탐색 및 정리	T 3. 창의적인 신규 방향 및 산출물 제작_text	T 4. 창의적인 신규 방향 및 산출물 제작_이미지
주 사용 프롬프트	1. instruction	○	○	○	○
	2. output indicator	○	○	○	○
	3. context		○	○	○
	4. role		○	○	○
	5. Few shot			○	○
예시		단순 정보 검색 및 단순 정보 정리, 요약	ex) 특정 양식에 맞춘 정보 검색 및 정보 정리ex) 특정 양식에 맞춘 교정 및 요약	특정 양식과 상황에 맞는 블로그 글 또는 광고 카피 제작	특정 양식과 상황에 맞는 이미지 및 도표 생성
프롬프트(입력 정보의 구조)의 복잡성		하	상	상	중
답변 교정을 위한 멀티턴의 수		하	중	상	상
산출될 기대 결과의 구체성		상	중	하	하
본 대화를 진행한 Heavy user interviewees		i-1, 2, 8, 10, 17, 21, 22	i-1, 2, 8, 10, 17, 21, 22	i-1, 2, 8, 10, 21, 22	i-1, 2, 17
본 대화를 진행한 Middle user interviewees		i-3, 7, 15, 18	i-3, 7, 18	i-3, 7, 15	i-3
본 대화를 진행한 Novice user interviewees		i-4, 9, 11, 12, 13, 14	i-,9, 11, 12, 13	i- 12, 13	

사전 설문 문항(Table 1. 5-1 ~ 5.5)을 통해 3순위까지의 LLM 서비스 주 사용 목적을 확인 후 해당 목적에 맞는 LLM 사용 화면 이미지를 인터뷰이들에게 요청하여 총 54가지의 대화 상황을 취합하였다. 이후 친화도를 기준으로 7가지 대화 목적(1. 정보검색 및 요약, 2. 정보 검색 및 정리, 3. 코딩 지원 및 보조, 4. 특정 기준에 기반한 작성 문서의 교정, 5. 아이디어 및 시나리오 제작, 6. 이미지 생성, 7. 키워드 기반 텍스트 작성)으로 수렴한 뒤 '타인에게 공유 되는가', '양식화되는가'의 여부와 '결과물이 객관적인 정리, 취합 등의 목적인가', '주관적이고 서비스의 창의성을 바탕으로 신규 산출물을 제작하는가'여부에 의해 4가지 대화 유형으로 나누었다. T1은 공유되지 않는 범위에서 자신을 위한 정보 탐색 및 정리 목적으로 사용하는 대화로서, 인터뷰이 전원이 사용하는 대화 유형이다. 단순 정보 검색 및 요약, 번역 등이 이에 해당하는데, 타 대화 유형에 비해 상대적으로 프롬프트의 복잡성이 낮고 멀티턴의 발생이 적은 대화로 볼 수 있다. T2의 경우 타인에게 보고 및 고유 목적으로 양식화된 정보 탐색 및 정리 텍스트가 이에 해당하는데, LLM 서비스에게 구체적인 작성 맥락 및 문체를 요구해야 하기 때문에 T1 대화에 비해 상대적으로 사용되는 프롬프트의 복잡성이 높아짐과 동시에 교정을 위한 대화의 턴수도 증가하게 된다. T3은 블로그 글 작성 및 광고 카피 작성 등과 같이 창의적이고 주관성이 높은 신규 텍스트를 제작하는 대화 유형으로, 프롬프트의 복잡성이 높아지고 멀티턴이 증가하는 양상을 보인다. 특히 장문의 T3 텍스트에 비해 단문의 T3 텍스트를 작성하는 경우 더욱 높은 멀티턴이 발생되는 경향이 관찰되었다. 전반적으로 사용자가 서비스를 통해 제공받을 결과가 다양하게 나타날 수 있으므로 지속적으로 멀티턴을 통해 개선안을 도출하고자 하는 경향을 확인하였다. T4의 경우, T3과 동일한 신규 산출물 제작의 대화 맥락이라 판단되지만, 텍스트가 아닌 이미지 생성에 목적을 둔 대화 유형으로서, 산출물의 성격은 T3과 상이하지만 Prompt 복잡성 및 멀티턴의 양 그리고 기대결과의 모호함 측면에서는 유사한 경향성이 확인되었다.

3. 3. 2. 집단별 프롬프트 사용량 분석

LLM 서비스의 대화 입력 구조인 Prompt에 대한 이해가 사용자들에게 상이한 대화 품질과 경험을 제공한다는 연구를 바탕으로(Pereira, 2023; Breanna et al.; 2024; Cho, 2024) 각 집단별 Prompt 사용 행태를 분석하고자 하였다. LLM에게 업무를 지시하는 방향의 프롬프트인 'Instruction(지시)', 대화 맥락에 대한 전반적인 정보 및 예시를 제공하는 'Few-shot(예시)'과 'Context(대화 맥락)', 산출된 대화의 제공 방식에 대한 기준을 제시하는 'Role(역할)'과 'Output Indicator(결과 제공 방식)'의 총 5개의 프롬프트 중 사용자가 Figure 3에 언급된 각 대화의 종류별로 몇 가지의 프롬프트를 사용하는지 분석하였다. Figure 5 좌측과 같이 각 대화 목적별로 사용한 프롬프트는 숫자 '1'로, 사용하지 않은 경우에는 숫자 '0'으로 코딩하여 엑셀을 통해 분석하였다. 분석 결과, 모든 집단에서 범용적으로 사용되는 T1과 함께 상대적으로 대화 결과의 복잡성이 높은 T2 대화와 T3 대화 또한, 서비스 사용 능력 및 TAM의 높음으로 인해 '업무 및 학습 맥락 내 LLM 서비스 사용 의존도'가 높을수록 Prompt 사용 양이 높은 것을 확인할 수 있었다.[Figure 5 우].

사분면 구분				1 사분면		
인터뷰이		i-1 박		i-6 김		
대화 레벨	대화목적	프롬프트 사용 유무		대화목적	프롬프트 사용 유무	
LV 1. 자신을 위한 정보 탐색	1. 코딩	input data	1	1. 정보요약	input data	1
		context	1		context	1
		role	0		role	1
		output indicato	0		output indicato	0
		instruction	1		instruction	1
		few shoot	0		few shoot	0
	2. 검색	input data	0	2. 단순정보 검색 및 요약	input data	0
		context	0		context	0
		role	0		role	1
		output indicato	1		output indicato	0
		instruction	1		instruction	1
		few shoot	0		few shoot	0
	3. 요약	input data	1	3. 업무용 정보 검색 및 요약	input data	0
		context	1		context	1
		role	0		role	1
		output indicato	1		output indicato	0
		instruction	1		instruction	1
		few shoot	0		few shoot	0
LV 2. 타인 양식화를 위한 정보 탐색	3. 업무에 활용할 텍스트 수정 교열	input data	1	발표논문 작성도움	input data	1
		context	1		context	1
		role	1		role	0
		output indicato	1		output indicato	0
		instruction	1		instruction	1
		few shoot	0		few shoot	1

* 프롬프트 사용 개수 누적 합산 (instruction 제외) / n=22

	T 1.	T 2.	T 3.	T 4.
노비스유저 그룹	16	6	12	2
미들 유저 그룹	10	2	7	0
헤비유저 그룹	35	28	32	3
총 합	61	36	51	5

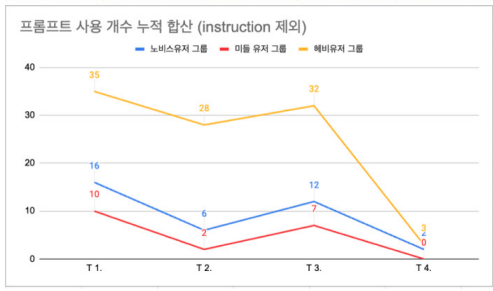

Figure 5 (Left) Interviewer Prompt Usage Analysis Form (Right) Prompt Usage Analysis

3. 3. 3. 동일 주제 내 대화창 개설 방식

Table 4 New Conversation Open Behavior by Analyzing the Number of Topics within the total number of Conversation

그룹	인터뷰이 번호	최근 순 25개 대화 내 주제의 수	주제당 사용하는 대화방 수	그룹별 평균
Heavy user group	i-1	25	1	1.62개
	i-6	20	1.25	
	i-8	18	1.38	
	i-10	14	1.78	
	i-22	11	2.27	
	i-21	12	2.08	
	i-2	16	1.56	
Middle User Group	i-3	22	1.13	1.21개
	i-15	20	1.25	
	i-17	25	1	
	i-18	17	1.47	
Novice user group	i-4	22	1.13	1.1개
	i-9	20	1.25	
	i-11	24	1.04	
	i-12	23	1.08	
	i-13	22	1.13	
	i-19	25	1	

LLM 서비스 좌측에 위치하여 사용자들의 대화창 개설 행태를 관찰할 수 있는 '히스토리' 분석을 통해 그룹 간의 주요 개설 기준에 대한 차이를 분석하고자 하였다. 이를 위해 최근 대화 순으로 25개의 대화를 총량으로 정의한 결과, 동일 주제에서 파생되거나 연결된 대화방의 비율이 그룹별로 상이하다는 점을 확인할 수 있었다. 해당 과정에서 총 대화내 주제의 개수가 적은 2명의 인터뷰이 [i-5 (Heavy user group), i-19 (Novice user group)]는 분석 대상에서 제외되었다. 단일 주제당 사용한 대화방 개수를 분석한 결과, Heavy user 그룹이 1.62개, Middle user 그룹이 1.2개, Novice user 그룹이 1.1개로 나타나 타깃 유저인 Heavy user 그룹이 상대적으로 큰 차이를 보이는 것을 확인할 수 있었다.

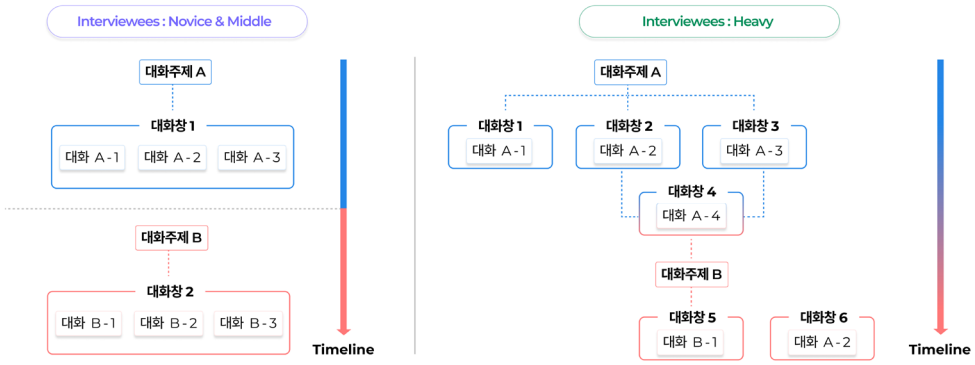

Figure 6 (Left) Initiative User's LLM Service Usage Pattern (Right) Heavy User's LLM Service Usage Pattern

즉, Novice user와 Middle user는 하나의 주제별로 하나의 대화창을 개설한다. 하나의 대화 주제에 속하는 여러 목적의 대화를 시도하는 경우, 추가적인 대화창 이동 없이 하나의 대화창에서 수행하는 패턴을 확인할 수 있었다. 때문에 각 대화 주제별로 분리된 히스토리 구조를 가지고 있다[Figure 6_좌]. 반면 Heavy user의 경우,

하나의 대화 주제에서 여러 대화 목적별로 신규 대화창을 개설한다. 각 대화창에서 획득한 정보를 요약하거나 병합하여 신규 대화창에 옮겨 작업하고, 대화창 내 특정 목적의 대화에서 새로운 대화 주제를 발굴하여 신규 대화창을 개설하는 등의 행태를 보였다[Figure 6_우]. 서비스 사용 능력과 수용 능력에 따른 서비스 의존도가 높은 Heavy user일수록 동일 대화 주제 내에서 개설하는 신규 대화창의 개수가 많으며, 대화창을 가지치기 하듯 유기적으로 사용하는 패턴을 발견할 수 있었다.

3. 3. 4. 사전 설문과 사용 화면 분석 인사이트

사전 설문과 LLM 사용 화면 분석을 통한 인사이트는 다음과 같다.

1) 본 연구 내 사용자 세분화의 정량적 기준인 '서비스 사용 능력'과 TAM(Technology Acceptance Model)의 측정값이 높을수록 업무 및 학습 맥락에서 LLM 서비스 사용 의존도가 높다.

2) '타인에게 공유되거나 양식화되는가'의 여부와, '대화 결과가 창의적이고 주관적인가'에 따라 대화는 총 4가지 유형으로 나뉘며, 각 대화 유형들은 프롬프트의 복잡성, 교정을 위한 멀티턴의 수, 서비스를 통해 산출될 기대 답변의 구체성과 모호성에 따라 상이하다는 특징을 가지고 있다.

3) 업무 및 학습 맥락에서 서비스 의존도가 높은 집단일수록 T1(자신을 위한 정보 탐색 및 정리), T2(타인, 양식화를 위한 정보 탐색 및 정리)와 T3(창의적인 신규 방향 및 산출물 제작_text) 대화 시 프롬프트 사용량이 많아지는 경향을 가지고 있다.

4) 업무 및 학습 맥락 내 서비스 의존도가 높은 집단일수록 동일한 대화 주제 내에서 개설하는 대화창의 개수가 많으며, 각 대화방 별로 유기적으로 연결되어 있다.

3. 4. In-depth 인터뷰 설계

앞선 인사이트의 풍부한 경험적인 이해를 위한 본 인터뷰의 목적은 총 2가지로 나뉜다.

첫 번째 목적은 인사이트 3번(프롬프트 사용량의 증가 추이)과 4번(대화창 개설을 바탕으로 한 히스토리 사용 행태) 기저의 멘탈 모델을 이해하고자 하며, 두 번째 목적은 4가지 대화 주제에 따라 사용자 그룹의 상이한 행태를 발견하고자 한다.

해당 2가지 목적을 바탕으로 비교군인 두 집단과 타깃 유저인 Heavy user 그룹간의 행태 차이 및 주요 인사이트를 얻고자, 2024. 01. 27.부터 4일간 총 20명을 대상으로 비대면 인터뷰를 진행하였고, 녹화된 인터뷰 영상을 엑셀 시트에 전사하여 분석하였다.

Figure 7 (Left) Conducting a Interview
(Right) Organizing and Transcribing Interview Utterances

4. 인터뷰 인사이트 기반 전략 도출

4. 1. 주요 행동 기저의 멘탈 모델 정의

각 대화 유형별 집단 간 행태 차이에는 인터뷰이들의 직업 및 사용 맥락이 상이하여 명확한 패턴을 확인할 수 없었으나, 사전 조사를 통해 발견한 프롬프트 사용량과 히스토리 대화창 개설 행태에서 집단별 차이점을 확인하여, Heavy user 그룹에서 관찰된 두 가지 주요 행동 기저의 멘탈 모델을 확인할 수 있었다.

첫 번째로, T2와 T3 대화에서 상대적으로 많은 양의 프롬프트를 사용한 Heavy user 그룹의 경우 U-1, 2, 4, 5와 같이 프롬프트 입력을 통해 만족스러운 품질의 대화 결과가 산출될 것이라는 기대와 함께, 장기적인 관점에서 높은 완성도의 프롬프트 대화 정보를 입력하여 대화방을 통한 만족도 또한 향상될 것이라 판단하였다.

반면 비교군인 두 집단의 경우, LLM 서비스를 통한 대화에 크게 의존 및 신뢰하지 않는 경향을 보이는데[U-7, 8], 이는 프롬프트에 대한 낮은 이해도를 가지고 장기간 서비스를 이용하여, 다수의 Hallucination을 경험한 것이 원인이라 판단된다.

두 번째로, 지속적으로 동일 대화 주제 내에서 신규 대화창을 개설하는 행태를 보인 Heavy user 그룹의 경우, U-10과 같이 대화가 장기화되는 경우, 특정 정보를 제외한 답변을 LLM 서비스 측에서 제공하거나, U-11, 12와 같이 높은 품질의 대화가 산출되었을 경우, 이후 추가적인 대화가 지속되었을 시 대화방 내 오염도가 증가하여 Hallucination이 발생할 것을 예상하여 지속적으로 신규 대화방을 개설하는 경향을 보였다. 이는 대화방에서 여러 정보의 병합을 최소화해야 높은 품질의 대화를 유지할 수 있을 것이라는 멘탈 모델로 인한 행동으로 보인다. 반대로 비교군인 두 집단의 경우 한 대화 주제 단위의 대화방 내에서 여러 대화 목적을 혼합해야 맥락 정보의 정확도가 높아진다고 믿거나[U-9, 11, 12], 대화방 내 오염도를 크게 신경 쓰지 않는 정도의 대화를 수행하는 모습을 보였다[U-4, 9, 18, 19].

Table 5 Define Mental Models Underlying Key Behaviors Through Interview Utterance Analysis

발견한 행동 및 사용 경험	인터뷰이 그룹	인터뷰이 번호	발화 내용(Utterance)	인사이트	인사이트 기저 멘탈 모델
업무 및 학습 맥락 내 서비스 의존도가 높은 집단일수록 T2, T3 대화 시 프롬프트 사용량이 많아진다.	Heavy User Group	i – 1, 5	U1. 인풋창 안에서 스스로 프롬프트 간의 연관성을 생각하며 교열하는 시간이 많은 편인데, 이렇게 하는 이유는 한 번에 좋은 답변을 얻을 수 있을 것이라는 생각 때문이야.	프롬프트의 사용을 통해 높은 품질의 대화가 가능함을 알기 때문에 대화 시 프롬프트의 구체성 및 복잡성이 높아진다.	상대적으로 Heavy user 집단은 LLM 서비스를 통한 대화 품질의 기대 수준이 높다.
		i – 1, 8, 22	U2. 구어체로 작성할 때보다, 프롬프트의 여러 역할을 이해하고 나니 훨씬 더 좋은 결과를 쉽고 빠르게 얻는 편이라, 귀찮더라도 꼭 작성하는 편이야.		
		i–2	U3. 블로그 글 작성 시에 어투와 예시를 넣는 것이 완성도를 높여주거나 수정이 적어지는 방향이라 생각되기 때문에, 역할(Role)과 예시(Few shot, context)를 필수적으로 입력하는 편이야.		
		i – 2, 5, 8	U4. 프롬프트를 통한 대화로 대화방에 다량의 정보가 입력되면, 그 대화방은 항상 좋은 답변을 주는 것을 알기 때문에 장기적인 측면에서 하나하나 대화에 시간이 들더라도 노력해요.		
		i – 1, 21	U5. 사람과 마찬가지로 논리적이고 이해하기 쉽게 설명해 줬을 때의 결과가 항상 만족스러운 부분이 있고, 프롬프트가 그 논리를 보강해주는 대화의 도구라고 생각해요.		
		i – 2, 8, 10	U6. 프롬프트를 통해서 대화하고 좋은 답변을 모으는 것이 익숙해지다 보니, 멀티턴이 발생하거나 수정을 지시할 때 구어체로 바뀌게 되는데 이 때 불안함이 생기는 거 같기도 해. 그래서 수정 시에도 프롬프트 구조를 유지하려 노력하는 것 같아.		
	Middle, Novice User Group	i – 4, 18	U7. 프롬프트라는 틀이 있는 건 아는데 잘 알지 못하고 , 대화를 가벼운 느낌으로 해나가는 것 같아서, 사용 후 가벼운 마음으로 제거하는 것 같고. 굳이 다시 찾아보거나 탐색하지 않는 것 같아.		
		i – 9	U8. 경쟁사 리서치 당시 틀린 정보들이 많은 탓에, 어차피 틀린 정보를 많이 알려주고 한두 개 건지는 거라고 생각해서 프롬프트까지 작성해 사용해야 하는지는 잘 모르겠어.		
		i – 11, 9, 17	U9. 특히 창의적인 텍스트를 작성할 경우의 LLM은 프롬프트의 유무보다 도메인에 대한 이해도가 낮아서 잘못된 대답을 한다고 생각하고, 프롬프트를 아무리 열심히 적는다고 해도 다른 정보를 끌어올 수 없다면 만족스러운 대답을 기대하기는 힘든 거 같아.		
업무 및 학습맥락 내 서비스 의존도가 높은 집단일수록 동일한 대화 주제 내에서 개설하는 대화창의 개수가 많다.	Heavy User Group	i – 1, 8	U10. 한 대화방 안에서 대화가 길어질 경우, LLM이 특정 내용을 기억하지 못한다고 생각될 때마다 같은 주제로 대화방을 신규 개설하게 돼.	대화방 내 정보의 병합 현상으로 발생하는 대화 오염도를 신경을 쓰기에 지속적으로 새 창을 열어 높은 품질의 대화를 유지하고자 한다.	
		i – 5, 6, 21	U11. 대화를 하다가 마음에 드는 대화가 나올 때마다 대화방을 옮기는 편이야. 이후의 대화를 통해 잘 정립된 정보 맥락이 무너질까봐 조심스럽거든. 그래서 그런지 대화방이 너무 많아지는 편인 것 같아.		
		i – 2, 6, 22	U12. 새로운 주제별로 대화창을 개설하려고 노력하지만, 바쁜 업무 중에는 분리가 어려워져 버리게 된 대화방도 여러 개인 것 같아서 아쉬울 때가 많았어. 다시 만족스러운 대화방을 만드는 건 매우 어려운데 말이지.		
		i – 6, 22	U13. 틈틈이 대화방을 옮김으로써 '깨끗한 대화'를 할 수 있는 것을 알고 나서부터는 노력해서라도 대화방을 옮기는 편인데, 어느 정도의 정보를 어떻게 이동하여 입력해야 되는지는 잘 몰라서 복사 붙여넣기를 하는 편인 거 같아.		
	Middle, Novice User Group	i – 9, 11, 12	U14. 대화창을 프로젝트 단위로 카테고리화해서 개설하여, 프로젝트 관련 대화는 그 대화창에서 주로 진행하는 것 같아.		
		i – 4, 18, 19	U15. 같은 대화 주제에서 꼬리 질문이나 목적이 생길 경우에는 혹시나 하는 마음에 본 대화창에서 대화를 이어나가는 편이야.		
		i – 3, 9, 13	U16. 일회성이 강한 대화 위주로 대화하다 보니 주제 안에서 여러 대화를 할 때도 있고 아닌 경우도 있어.		

4. 2. 주요 행동별 페인포인트 도출 및 UX 전략 제안

Figure 8_ P1. LLM 대화에 대한 높은 기대 품질로 인해, 대화방 내 오염도를 고려하여 동일 주제 내에서 지속적으로 신규 대화방을 생성하는 사용자들은 완성된 답변을 복사하여 타 대화방으로 이동하거나, 2개 이상의 높은 품질의 답변을 요약 및 병합하여 신규 대화방으로 이동하는 행동을 보인다. 이 때 복사하거나 병합한 답변에 앞뒤 대화 맥락이 포함되지 않아 이동한 대화방 내 품질이 떨어지는 문제가 발생함을 확인하였다.

Figure 8_ P2. 대화 품질에 높은 기대로 프롬프트의 복잡성이 높아지는 사용자는, 일정 기간 이후 서비스에 재유입하여 대화를 재탐색하는 경우가 빈번함을 확인하였다. 이 때, 같은 주제로 생성한 여러 대화방 중 대화를 재탐색하는 과정에 어려움이 발생한다. 재탐색에 실패할 경우 다시 복잡한 프롬프트를 입력해야 한다.

Figure 8_ P3. 프롬프트를 통한 구조적 대화에 익숙한 사용자들은 산출된 결과에 대한 수정 지시와 같은 멀티턴이 발생될 때, 구어체 및 단문의 형식으로 프롬프트 구조가 해체되는 경향을 보이는데, 이때 Heavy user 그룹의 사용자 경우 프롬프트 구조가 아닌 단문 및 구어체로 수정 요청할 경우, 정확한 정보 입력이 진행되었는지 여부에 대한 불안감을 가지고 있음이 발견되었다.

Figure 8_ P4. T2와 T3의 대화 중 평소에 해보지 못한 익숙하지 않은 목적의 대화를 수행하는 경우에도 자신이 알고 있는 구조의 프롬프트만을 입력하는 경향이 발견되었다. 이때 추가해야 하는 프롬프트의 종류, 작성 정보의 양에 대한 명확한 기준이 없음에 불안감을 느끼는 문제가 발생한다.

서비스 내 주요 행동을 통해 정의한 4가지 페인포인트를 바탕으로, 개선 경험에 대한 3가지 UX 전략 방향을 총 7가지 인터페이스를 통해 구체화하였다.
P1과 P2에 대한 개선으로 두 개의 기능으로 구성된 전략 1을 제안하였다. 대화 주제별 히스토리의 위계 구조를 제안하여 대화 탐색을 용이하게 하였다[Figure 9. UX1-1]. 또한 대화에 필요한 답변의 범위를 선택할 수 있게 하여 대화 오염도 이슈를 최소화하는 방향의 기능을 제공하였다.[Figure 10. UX1-2]. 이를 통해 대화 탐색 및 답변 제작 경험을 개선하는 방향의 전략이다.
P2에 개선을 위해 총 3개의 기능으로 구성된 전략 2를 제안하였다. 만족스러운 대화를 저장할 경우 앞뒤 대화 맥락을 선택하여 저장하고[Figure 11. UX2-1], 저장한 대화끼리 결합하여[Figure 11. UX2-2] 해당 정보가 포함된 신규 대화방을 개설[Figure 12. UX 2-3]하는 기능의 제공의 통해 답변 정보의 대화방 내, 대화방 간 이동을 용이하게 하는 경험을 제공하였다.
P3과 P4에 대한 개선의 경우 2가지 기능으로 제안되었는데, 글의 목적에 따라 필요한 프롬프트의 구조와 완료율 데이터를 제공하고[Figure 13. 3-1], 입력한 프롬프트 구조를 통해 답변을 수정하는 기능[Figure 14. 3-2]을 제공하여 프롬프트 기반 대화 경험을 개선하였다.

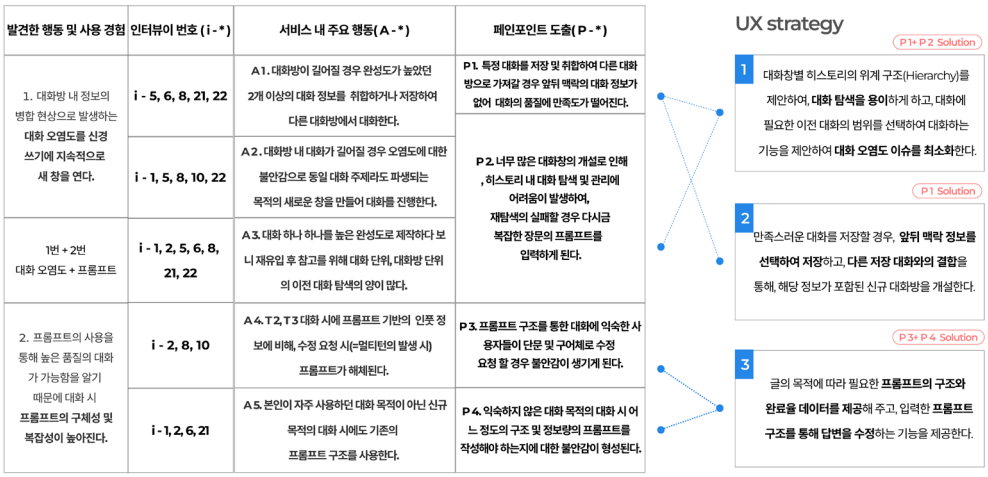

Figure 8 Define Pain points and Set Strategic Direction Based on Key Behaviors in the LLM Service

5. 전략별 디자인 제안

5. 1. 히스토리의 Hierarchy 제안 및 대화 정보 범위 선정

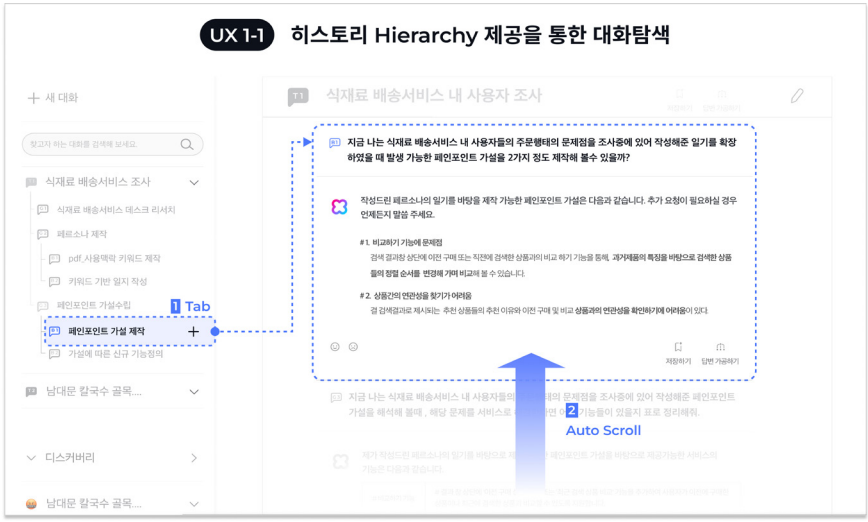

Figure 9 Improve Browsing Experience by Providing Hierarchy in Conversation History

Figure 10 Minimize Contamination by Selecting Dialogue Information

1번 전략은 대화창에서 각 대화를 주제 〉 목적 〉 소목적 단위로 분류하고, 인터페이스를 통해 대화 계층을 구조화하여 이전 대화 탐색을 용이하게 한다. 동시에 대화방 내에서 필요한 답변의 범위를 지정하여 대화하거나, 해당 범위의 대화를 바탕으로 새 대화창을 생성하여 대화방 내에서 타 정보로 인한 답변 오염도를 최소화하는 전략이다. 좌측 인터페이스에서 제공하는 대화의 계층 구조에서 과거 답변을 선택할 경우, 화면이 해당 답변으로 스크롤되어 이동한다. 이는 대화방 내에서 각 답변의 탐색을 수월하게 한다. 계층 구조는 입력창에서도 확인 가능하며, 입력창에서 각 대화 목적별 배지 컴포넌트를 탭하면 '선택한 정보로 대화하기' 버튼이 활성화되는 기능으로 나타난다. 해당 기능은 대화방 내 일부 답변 정보만 포함시켜 대화를 진행하여 병합된 여러 대화 정보들로 인해 발생하는 Hallucination을 최소화한다. 또한, 목적별 배지 컴포넌트를 탭하여 '선택한 대화로 새 대화창 열기'를 실행한 경우, 해당 맥락 정보를 포함한 새 대화창을 개설할 수 있다. text 간의 유사도를 기반으로 대화방 내 답변 정보의 계층 구조를 제공하는 것을 기반으로 하는 전략 1의 경우, Text embeding(=Vector)을 추출하고 각 Vector 간에 Cosine Similarity를 계산해서 유사도를 측정한다는 방향의 기술을 통해 구현 가능성을 높일 수 있을 것이라 판단된다(TechClaw, 2023).

5. 2. 맥락정보를 포함한 대화 저장하기 및 가공하기

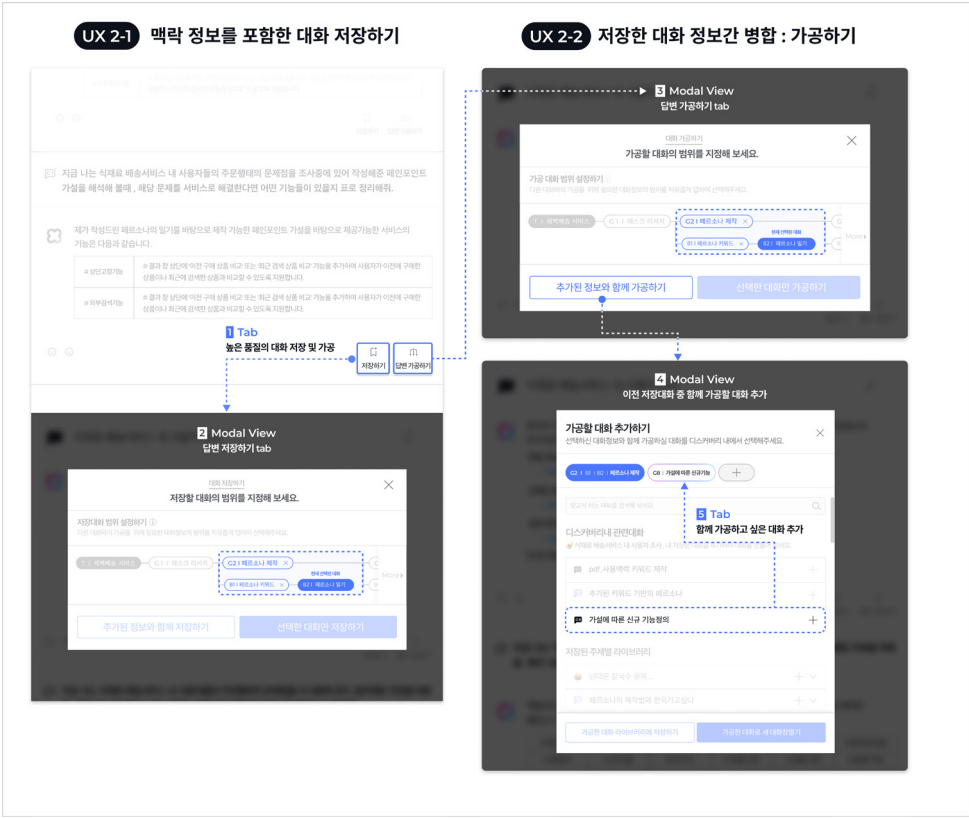

Figure 11 Ability to Save and Combine answers with Contextual Information

본 전략은 만족스러운 품질의 답변을 저장하거나 타 대화방으로 복사하여 이동하는 경우, 앞뒤 맥락 정보가 없어서 발생하는 대화 품질의 저하를 개선하고자 한다. 두 개 이상의 답변 정보를 취합하거나 가공하여 해당 답변의 정보가 포함된 새 대화창을 제작하고, 높은 품질의 대화 정보를 타 대화방으로 용이하게 이동시킬 때 사용할 수 있는 기능이다. 만족스러운 답변 생성 시, 우측 하단의 '저장하기' 기능을 탭하면 선택한 대화를 중심으로 앞뒤에 함께 저장하고자 하는 대화 맥락을 선택할 수 있는 인터페이스가 제공된다. 이후 생성된 또

다른 답변과 이전에 저장된 대화 정보를 병합하고자 할 경우, 답변 인터페이스 우측 하단의 '답변 가공하기'를 탭한다. 제공된 모달 뷰를 통해 이전에 저장한 대화를 선택하면, 2가지 대화 정보를 병합할 수 있으며, 단순 병합 후 저장 기능을 제공할 뿐만 아니라 해당 병합 대화 정보를 내포하는 신규 대화창을 개설할 수 있는 기능이 제공된다. 대화 정보간의 병합(=가공하기) 및 맥락 정보와 본 대화 간의 병합(=저장하기)이 중요한 전략 2의 경우, 대화 간의 가중치를 둘 수 있는 데이터베이스(DB) Schema를 구축하여 DB 테이블을 관리, 이 후 쿼리 기능을 통해 신규 대화방으로 정보를 불러오는 방향의 기술로 구현이 가능할 것이다(Yuhuan, 2023).

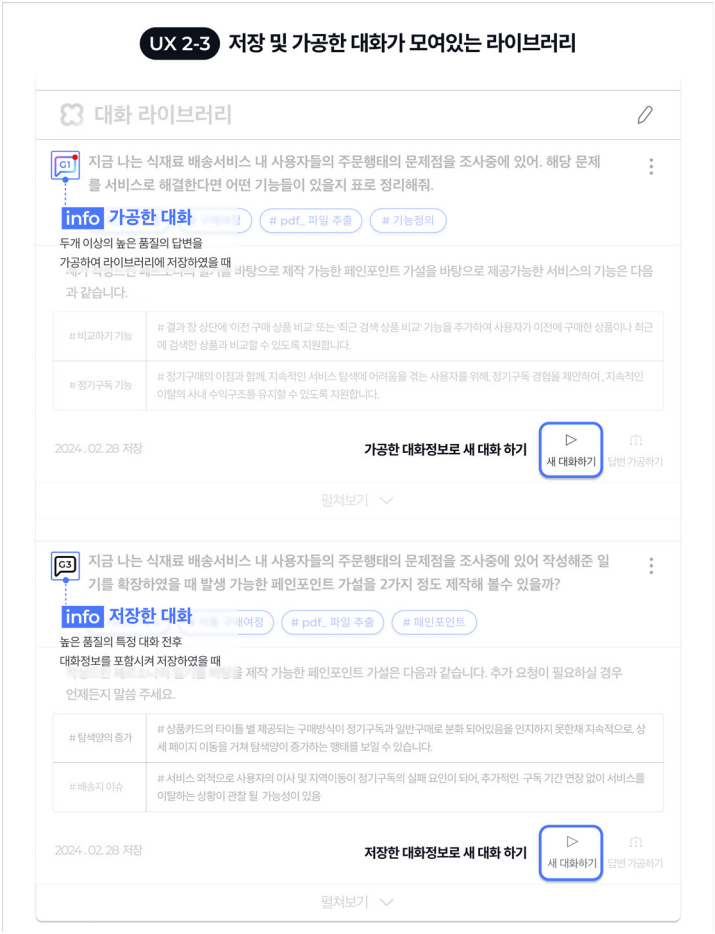

Figure 12 Saved and Combined Conversation Libraries

5. 3. 프롬프트 가이드라인 제공 및 답변 수정 기능 제안

본 전략은 목적별로 세분화된 프롬프트의 구조와 작성한 텍스트 정보의 완성도 지표를 제공하여, 대화 작성 시의 기준을 제공한다. 사용자의 불안감을 최소화하고, 구어체와 단문으로 해체되지 않은 프롬프트를 유지하며 제공된 답변을 수정하는 방향의 전략이다.

원하는 대화 목적을 선택할 경우 해당 목적에 필요한 프롬프트 구조와 높은 완성도에 답변이 산출될 가능성을 정량적으로 제시하는 작성 완료율 데이터를 퍼센트로 좌측 상단에 제공한다.

Figure 13 Provide Prompting Guidelines and Completion Percentage per Conversation Purpose

산출된 답변을 수정할 경우 '답변 수정하기' 버튼을 탭하게 되면 이전에 작성된 프롬프트 구조에 추가되면 좋은 정보의 양 및 종류에 대한 큐레이션을 제공하고, 동시에 '하이라이트' 기능을 제공한다. '하이라이트' 기능이란, 제공된 답변 중 조금 더 부각되거나 강조 되었으면 하는 문장 및 키워드를 선택하여, 해당 키워드와 문장을 바탕으로 답변을 수정하는 기능이다. 프롬프트 구조 기반의 정보 입력 수정 기능의 제공과 하이라이트 기능을 통해 T2와 T3 대화의 진행 시 발생하는 불안감을 최소화 할 수 있는 기능이라 판단된다. 해당 기능은 서비스 내 축적된 사용자들의 목적별 대화 데이터를 활용한 입력필드를 구체화하고 세분화하여 구현 가능할 것이다.

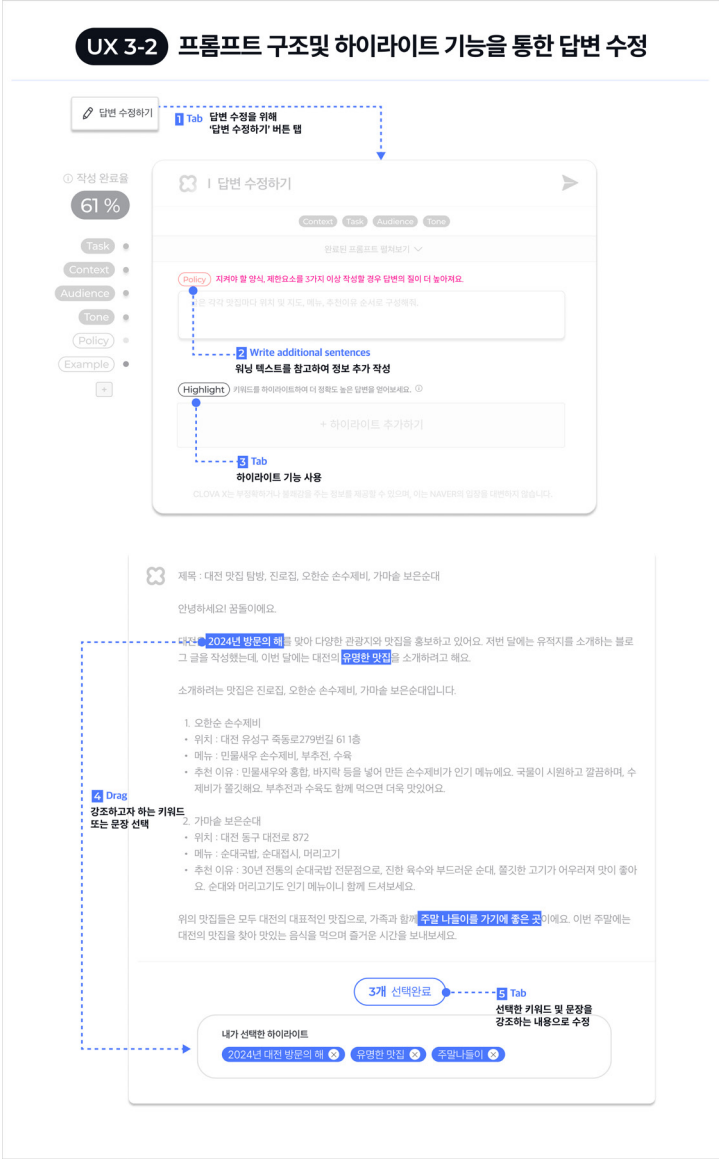

Figure 14 Prompt-Based Answer Editing and Highlighting Features

6. 사용성 테스트

6. 1. 사용성 테스트 설계

제안된 3개의 전략을 통한 LLM 서비스 사용 경험의 개선 정도를 확인하기 위해 사용성 테스트를 2024년 3월 11일부터 약 5일간, 타깃 집단인 Heavy user 그룹 13명, 비교군인 Middle user 그룹 8명과 Novice user 그룹 12명, 총 33명을 대상으로 진행하였다. 테스트는 하나의 세션당 3명에서 8명 사이 인원으로 총 12개의 세션으로 진행되었고 이 중 5개 세션 그룹은 비대면, 7개의 세션 그룹은 대면으로 진행되었다. 이는 타임 리소스를 최소화하면서도 전체 테스트 이후 세션 그룹별 사후 인터뷰를 통해 추가적인 개선 방향에 대한 인사이트를 얻기 위해서였다. 사용성 테스트는 총 5개의 과정으로 구성되었다. 1) 각 전략에 대응되는 기존 경험 Task를 약 5분간 진행한 뒤 2) PU(지각된 유용성)와 PEU(지각된 사용 용이성)로 이루어진 TAM

평가 척도를 측정하였다. 3) 이후 개선 전략 경험에 대한 프로토타입 구동 영상 및 설명을 청취 후 간단한 질의응답을 퍼실리테이터와 진행하였으며 피실험자 간에 대화는 통제하여 진행하였고 4) TAM 평가 척도를 통해 평가하였다. 5) 실험이 종료된 이후 그룹별 테스트 참가자들 간의 의견을 주고받는 사후 인터뷰를 통해 추가적인 인사이트를 얻고자 하였다.

최종적으로 취합된 전후 경험에 대한 PU, PEU 데이터를 바탕으로 비교군을 포함한 전체 실험 대상자(N=33)는 대응 표본 t-test를, 타깃 유저인 헤비 유저(N=9)를 대상으로 윌콕슨 부호-순위 검정(Wilcoxon Signed-Ranks Test)을 수행하여 전략별 개선 정도를 통계적 유의성을 통해 검증하고자 하였다. 이때 비대면, 대면에 따른 PU의 평균 차이는 유의하지 않았으며(비대면 그룹 M=25.9, SD=8.28 / 대면 그룹 M=28, SD=7.56 / t=0.730, p=0.472), PEU의 평균 차이도 유의하지 않았다(비대면 그룹 M=25.84, SD=7.48 / 대면 그룹 M=27.25, SD=8.28 / t=0.505, p=0.618). 따라서 실험 진행 조건에 따른 영향은 없었다고 볼 수 있다.

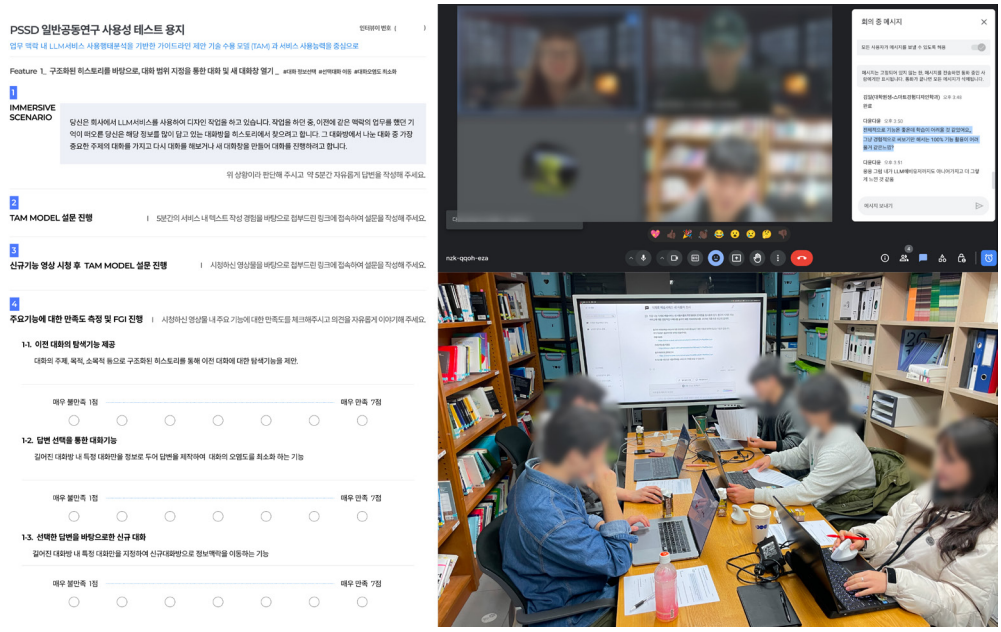

Figure 15 (Left) Usability Test Form (Right) Usability Test Scenes

6. 2. 통계검정

Table 6 Verifying the Enhancement Experience with a Paired Sample T-test

n=33

평가항목		기존 LLM		개선 LLM		T-test	
		M	SD	M	SD	T	p
경험 1.이전 대화 탐색 및대화 오염도 조절	PU	26.00	6.964	35.18	3.762	-7.147	0.001***
	PEU	17.36	4.968	20.12	4.872	-2.448	0.02*
경험 2.특정 대화 저장 및타 대화방으로 이동	PU	24.42	8.239	32.27	6.022	5.107	0.001***
	PEU	16.45	6.093	18.09	5.714	-1.083	0.287
경험 3.프롬프트 기반 작성및 수정	PU	25.58	7.176	36.52	3.850	-8.747	0.001***
	PEU	18.67	4.728	21.79	4.159	-2.625	0.013*

*p<.05 **p<0.1 ***p<.001

3가지 개선 전략과 기존 경험 간의 PU(인지된 유용성), PEU(인지된 사용 용이성)의 차이를 분석하기 위해 비교군을 포함한 전체 실험대상자(N=33)의 데이터를 기반으로 대응표본 t-test를 실시하였다. 1번과 3번 전략의 경우 PU는 0.001 기준으로, PEU는 유의수준 0.05를 기준으로 통계적 유의함을 보여, 기존 경험에 비해 유용성과 사용 용이성에서 개선됨이 밝혀졌다.

1번 전략은 좌측 히스토리 및 입력창을 통해 제공되는 구조도를 통해 특정 답변을 탐색하면, 시간 단축 및 대화방 내 대화 구조를 조금 더 명확히 할 수 있을 것이라는 의견을 얻을 수 있었다. 또한 특정 높은 품질의 답변 정보가 내포되어 있는 신규 대화창을 개설할 경우, 특정 대화를 복사하여 신규 대화창에 붙여넣는 과정에서 발생하는 대화 품질의 저하를 최소화할 수 있을 것이라는 긍정적인 의견을 사후 인터뷰를 통해 얻을 수 있었다.

3번 전략은, 사후 인터뷰를 통해 테스트 참가자 그룹에 상관없이 가장 높은 만족도를 보였는데, 프롬프트 작성 시에 제공되는 완료율 데이터 및 하이라이트 기능을 통해 작성 및 수정 시 발생하는 불안감을 해소할 수 있을 것이라는 의견이 지배적이었다.

2번 전략의 경우 PU는 유의 수준 0.001을 기준으로 통계적 유의함을 보였지만 PEU의 경우 유의 수준 0.287로 유의성을 보이지 않았는데, 해당 PEU의 원인을 사용성 테스트 대상자 표집에서 확인할 수 있었다. 이는 일부의 타깃 유저를 제외한 표집 대상자들의 경우 특정 대화의 저장 및 가공을 통해 타 대화방으로 정보를 이동시키는 전략 2에 대한 기존 경험이 없었기 때문에, 두 경험 간의 PEU(사용 용이성) 차이를 느낄 수 없었음 을 사후 인터뷰를 통해 확인할 수 있었다.

이후 표집 대상자들의 사전 설문 데이터를 기반으로 Sorting된 헤비유저 9명을 대상으로 윌콕슨 부호-순위 검정(Wilcoxon Signed-Ranks Test)을 수행하였다. 전략 1은 'PU(Z= 2.67, p 〈 .008)', 'PEU(Z= 1.332, p 〈 .183)', 전략 2는 'PU(Z= 2.298, p 〈 .022)', 'PEU(Z= 1.021, p 〈 .307)', 마지막으로 전략 3은 'PU(Z= 2.805, p 〈 .0005)', 'PEU(Z= 1.684, p 〈 .092)',로 도출되어, 3가지 전략 모두 유용성에서는 개선됨을 확인 할 수 있었지만, 사용 용이성의 경우 전후 경험 간개선의 유의함을 확인할 수 없었다.

사후 인터뷰를 통해, 개선 전략 자극물을 직접 구동해 보지 못하고 영상 및 설명으로 대체됨으로 인한 피실험자의 능동성 결여가 학습 용이성이 유의하지 않음에 대한 상위 원인임이 확인되었는데, 이는 전체 12개의 실험 세션 중 직접 구동이 가능한 대면과 불가능한 비대면 실험 세션 간의 변인통제를 위함으로 불가피하였다 판단된다.
추가적으로 전략 1의 경우, 출력 정보 오염도를 최소화한 '선택 정보로 대화하기' 기능 내 제공되는 대화 구조도의 badge 인터페이스가 Touchable 컴포넌트인지에 대한 인식이 어려웠다는 점과, 완료된 대화 정보를 선택 후 추가로 대화하는 여정 내 가이드 제공이 부족하다는 의견이 취합되었다.
전략 2는, 신규로 제작된 대화 정보와 이전에 사용자가 저장해 둔 대화 정보를 합쳐 가공하는 기능에 있어서, 기존에 대화 정보가 저장되어 있는 히스토리에 대한 언급 및 설명이 실험 시 미비했다는 것을 확인할 수 있었고, 전략 3의 경우 프롬프트 기반 대화 시 높은 품질의 대화 정보 제작을 위한 완료율 퍼센트 제공 측면에서 기술 구현 가능성에 대한 신뢰도 측면의 이슈로 인해 실험 과정 전반에 발생한 편향이 사용 용이성 판단 시에 발생하였다는 점이 확인되었다. -이를 통해, 신규 기능에 대한 사용 여정 내 tooltip과 같은 가이드 제공을 통한 지속적인 사용 경험의 숙성이 필요하다 판단되었다.

5. 결론 및 논의

본 연구는 서비스 사용 능력과 TAM(Technology Acceptance Model)에 따라 세분화된 사용자 유형별 LLM(Large Language Model) 서비스 내 주요 사용자 행태를 분석하여 신규 UX 전략을 발굴 및 검증하고자 하였다. 이를 위해 선행 연구를 기반으로 한 2가지 사용자 세분화 척도를 바탕으로, Positioning Map에 정량적으로 배치하여 사용자들을 유형화하였고, 유형별 서비스 내 주요 행태에 대한 분석을 각 사용자별 서비스 사용 화면 이미지를 취합하여, 좌측 대화 히스토리 사용 행태와 프롬프트 사용 행태를 기준으로 분석하였다. 이후 발견된 패턴 및 인사이트 기반의 멘탈 모델을 이해하기 위해 in-depth 인터뷰를 통해 개선 전략을 도출하였고, 해당 전략의 시각화 및 사용성 테스트를 통해 유용성 및 학습 용이성의 개선 정도를 정성 및 정량적 방법으로 검증하였다. 연구 결과, 사용자의 서비스 내 주요 행동과 페인포인트를 기반으로 제작된 전략 1번과 3번의 경우, TAM 평가 척도인 PU와 PEU 모두에서 기존 서비스 경험에 비해 개선되었음을 통계적으로 검증하였다. 주요 행동과 페인포인트를 바탕으로 제작된 전략 2의 경우, PU의 개선 정도는 통계적으로 검증되었지만, PEU의 경우 검증하지 못하였는데, 이는 사용성 테스트 시의 표집의 한계 및 실험의 변인 통제를 목적으로 한 개선안의 영상 제공 등이 주요 요인인 것으로 사후 인터뷰를 통해 확인하였다. 본 연구는 다양한 업무 맥락 내에서 활용되는 LLM 서비스의 사용자별 주요 행태와 패턴을 다각도로 분석 및 정의하여, 현재 LLM 연구 도메인 내에서 미비하다 판단되는 LLM 인터페이스 경험 연구를 수행한 데에 의의가 있다. 또한 도출된 주요 인사이트를 User Interface 레벨로 구체화하여 사용성 테스트 및 검증의 과정을 통해 실증적인 연구를 제공했다는 점 또한 유의미한 의의라 판단된다.

하지만 3-2-1에서 제안한 정량적 Positioning Map 내 인터뷰이들을 그룹화한 기준 문항인 '서비스 및 학습 맥락 내 LLM 서비스 의존도' 이외의 다양한 경험 요인에 대한 발굴을 통해 새로운 그룹 세분화의 기준이 수립될 수 있음이 연구적 한계라 판단된다. 또한 사용성 테스트 당시 개선안에 대한 제공 방식이 피실험자들의 직접 구동이 아닌 구동 시연 및 설명이었기에, 평가 척도인 PEU에 대한 능동적 체험이 불가능했다는 점과 함께 신규 기능으로 인한 복잡도로 인해 사용 방법에 대한 추가적인 가이드라인의 제공이 필요하다는 점 또한 한계로서 존재한다. 또한 전략 도출의 바탕이 된 인터뷰이와 사용성 테스트 대상자 대부분이 겹침으로 인해, 두 가지 실험 만족도 평가에 오염이 발생하였을 가능성 또한 고려해야 할 부분이라 판단된다. 향후 테스트 자극물과 실험 상황의 개선을 통해 더욱 더 정확한 검증 방향 및 제안과, 서비스 내 사용자 세분화 기준에 대한 추가적인 경험 요인의 발굴을 통해 연구를 고도화할 수 있을 것이다.

References

1. Acharya, A., Singh, B., & Onoe, N. (2023). LLM Based Generation of Item-Description for Recommendation System. *RecSys '23 Proceedings of the 17th ACM Conference on Recommender Systems*, 1204-1207.
2. Agossah, A., Krupa, F., Silva, M. P. D., & Callet, P. L. (2023). LLM-Based Interaction for Content Generation: A Case Study on the Perception of Employees in an IT Department. *IMX '23: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 237-241.
3. Ahmed, T., & Devanbu, P. (2023). Few-shot training LLMs for project-specific code-summarization. *ASE '22: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, 177*(1), 1-5.
4. Alaofi, M., Gallagher, L., Sanderson, M., Scholer, F., & Thomas, P. (2023). Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1869-1873.
5. Ayoobi, N., Shahriar, S., & Mukherjee, A. (2023). The Looming Threat of Fake and LLM-generated LinkedIn Profiles: Challenges and Opportunities for Detection and Prevention. *HT '23: Proceedings of the 34th ACM Conference on Hypertext and Social Media, 38*(1), 1-10.

6. Chen, B., Mustakin, N., Hoang, A., Fuad, S., & Wong, D. (2023). VSCuda: LLM based CUDA extension for Visual Studio Code. *SC-W '23: Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing*, 11-17.
7. Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., & Xiao, Y. (2023). Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. *CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 245-255.
8. Cho, S. Y. (2024). Generative Artificial Intelligence and Narrative Creation : A Focus on Prompts. *시민인문학[Citizen and Humanities], 46*(7), 189-213.
9. Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *IT Usefulness and Ease of Use, 13*(3), 319-340.
10. Duan, P., Warner, J., & Hartmann, B. (2023). Towards Generating UI Design Feedback with LLMs. *UIST '23 Adjunct: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 70*(1), 1-3.
11. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *arXiv:2303.10130v5* [econ.GN] 21 Aug 2023 1OpenAI 2OpenResearch 3University of Pennsylvania August, https://arxiv.org/abs/2303.10130
12. Fede, G. D., Rocchesso, D., Dow, S. P., & Andolina, S. (2023). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. *C&C '22: Proceedings of the 14th Conference on Creativity and Cognition*, 623-627.
13. Hansen, W. J. (1972). User engineering principles for interactive systems. *Proceedings of the 7th ACM International Symposium on Pervasive Displays,* 523-532.
14. Hausi A, M., Litoiu, M., Rivera, L. F., Rasolroveicy, M., Villegas, N. M., Tamura, G., ... & Shwartz, L. (2023, September). Proactive Continuous Operations using Large Language Models (LLMs) and AIOps. In Proceedings of the 33rd Annual International Conference on Computer Science and Software Engineering (pp. 198-199).
15. Jaimes, A. (2023). Multimodal AI & LLMs for Peacekeeping and Emergency Response. *MM '23: Proceedings of the 31st ACM International Conference on Multimedia*, 3-4.
16. Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Reilly, A. L. (2024). Evaluating LLM-generated Worked Examples in an Introductory Programming Course. *ACE '24: Proceedings of the 26th Australasian Computing Education Conference*, 77-86.
17. Karolus, J., & Schmidt, A. (2018, June). Proficiency-Aware Systems: Adapting to the User's Skills and Expertise. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays* (pp. 1-2).
18. Kim, C. J. (2017). A Study on the Characteristics of Mobile on TAM. *유통경영학회지[Korea Research Academy of Distribution and Management Review], 20*(3), 115-126.
19. Kulkarni, C., Wu, T., Holstein, K., Liao, Q. V., Lee, M. K., Lee, M. N., & Subramonyam, H. (2023). LLMs and the Infrastructure of CSCW, *CSCW '23 Companion: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, 408-410.
20. Lewis, J. R. (2019). Comparison of Four TAM Item Formats: Effect of Response Option Labels and Order. *Journal of Usability Studies, 14*(4), 224-236.
21. Li, A., Wu, J., Bigham, J. P. (2023). Using LLMs to Customize the UI of Webpages. *UIST '23 Adjunct: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 45*(1), 1-3.
22. Li, L., Zhang, Y., Chen, L. (2023). Prompt Distillation for Efficient LLM-based Recommendation. *CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1348-1357.
23. Lu, Y., Deng, B., Yu, W., & Yang, D. (2023). In *MM '23: Proceedings of the 31st ACM International Conference on Multimedia*, 4053-4064.
24. Ma, R., Wang, J., Qi, Q., Yang, X., Sun, H., Zhuang, Z., & Liao, J. (2023). Poster: PipeLLM: Pipeline LLM Inference on Heterogeneous Devices with Sequence Slicing. *ACM SIGCOMM '23: Proceedings of the ACM SIGCOMM 2023 Conference*, 1126-1128.

25. Mondal, R., Tang, A., Beckett, R., Millstein, T., & Varghese, G. (2023). What do LLMs need to Synthesize Correct Router Configurations?. *HotNets '23: Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, 189–195.
26. Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., & Pengo, M. F. (2023). Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. *GoodIT '23: Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 205–212.
27. Morbidoni, C. (2023). Poster: LLMs for online customer reviews analysis: oracles or tools? Experiments with GPT 3.5. *CHItaly '23: Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, 38*(1), 1–4.
28. Okita, T., Ukita, K., Matsuishi, K., Kagiyama, M., Hirata, K., & Miyazaki, A. (2023). Towards LLMs for Sensor Data: Multi-Task Self-Supervised Learning. *UbiComp/ISWC '23 Adjunct: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, 499–504.
29. Park, J., & Choi, D. E. (2023). AudiLens: Configurable LLM-Generated Audiences for Public Speech Practice. *UIST '23 Adjunct: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 122*(1), 1–3.
30. Pereira, J. D. Z., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can&t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 437*(1), 1–21.
31. Rastogi, C., Ribeiro, M. T., King, N., Nori, H., & Amershi, S. (2023). Supporting Human-AI Collaboration in Auditing LLMs with LLMs. *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 913–926.
32. Rodríguez-de-Vera, J. M., Villacorta, P., Estepa, I. G., Bolaños, M., Sarasúa, I., Nagarajan, B., &Radeva, P. (2023). Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition. *MADiMa '23: Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, 43–52.
33. Rossetto, F., Dalton, J., & Smith, R. M. (2023). Generating Multimodal Augmentations with LLMs from Song Metadata for Music Information Retrieval. *LGM3A '23: Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, 51–59.
34. Sarabi, A., Yin, T., & Liu, M. (2023). An LLM-based Framework for Fingerprinting Internet-connected Devices. *IMC '23: Proceedings of the 2023 ACM on Internet Measurement Conference*, 478–484.
35. TechClaw. (2024, July 11). *Cosine similarity between two arrays for word embeddings*[Web log post]. Retrieved from https://medium.com/@techclaw/cosine-similarity-between-two-arrays-for-word-embeddings-c8c1c98811b
36. Tsai, W. E., & Liu, Y. C (2023). Aisen – Web-Based Gaze-Tracking Assistive Communication Interface with Word Cards Generated by LLMs. *UIST '23 Adjunct: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 116*(1), 1–3.
37. Ye, Q., Xu, H., Yan, M., Zhao, C., Wang, J., Yang, X., Zhang, J., Huang, F., Sang, J., & Xu, C. (2023). mPLUG-Octopus: The Versatile Assistant Empowered by A Modularized End-to-End Multimodal LLM. *MM '23: Proceedings of the 31st ACM International Conference on Multimedia*, 9365–9367.
38. Yi, S. H., & Kim, S. I. (2023). A Comparative Study on LLM-based AI Service User Experience – Focusing on ChatGPT and Clova X. *상품문화디자인학연구[Journal of Cultural Product & Design], 75(2), 11–22.*
39. Yin, B., Xie, J., Qin, Y., Ding, Z., Feng, Z., Li, X., & Lin, W. (2023). Heterogeneous Knowledge Fusion: A Novel Approach for Personalized Recommendation via LLM. *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*, 599–601.
40. Yoon, J. Y., & Shin, Y. (2023). Large Language Models are Few-shot Testers: Exploring LLM-based General Bug Reproduction. *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2312–2323.

41. Zhang, P., Jaipersaud, B., Ba, J., Petersen, A., Zhang, L., & Zhang, M. R. (2023). Classifying Course Discussion Board Questions using LLMs. *ITiCSE 2023: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 2*, 658.
42. Zhao, Z., Song, S., Duah, B., Macbeth J., Carter, S., Van, M. P., Bravo, N. S., Klenk, M., & Sick, K. (2023). More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. *C&C '23: Proceedings of the 15th Conference on Creativity and Cognition*, 368-370.

LLM 서비스 내 헤비 유저 사용 행태 분석에 기반한 인터페이스 제안

진종현[1], 이소영[1], 박채린[1], 연명흠[2*]

[1]스마트경험디자인학과, 학생, 국민대학교, 서울, 대한민국

[2]스마트경험디자인학과, 교수, 국민대학교, 서울, 대한민국

초록

연구배경 LLM(Large Language Model)은 논리적 추론, 질문 응답, 코드 생성 등의 다양한 산업 전반에서 활용이 확대되고 있다. 하지만 LLM 연구의 전반적인 방향성 및 분석 대상은 서비스 시스템을 통한 가용성 및 효율성에 치중되어 있으며, 서비스 내 사용자에 대한 행태 분석에 기반한 인터페이스 연구는 미비한 실정이다.

연구방법 본 연구는 LLM 서비스 내 사용자 집단별 상이한 이용 행태 기저의 멘탈 모델에 대한 정의를 목표로, 사전 설문 데이터를 통한 정량적 Positioning Map을 제안하여 사용자를 3집단으로 정의한 뒤, 서비스 사용 화면 분석 및 in-depths 인터뷰를 통해 개선 전략을 제안 및 검증하였다.

연구결과 조사를 통해 정의한 3집단 중 타깃 집단인 Heavy 유저 그룹의 경우 비교군인 Middle, Novice 유저에 비해 LLM 서비스를 통한 대화 품질의 기대 수준이 높다는 기저의 멘탈 모델을, 주요 행동인 프롬프트 사용 행태 및 신규 대화창 개설 행태 등을 통해 정의하였고 이를 바탕으로 3가지 개선 전략(1. 히스토리의 Hierarchy 제안 및 대화 정보 범위 선정, 2. 맥락 정보를 포함한 대화 저장하기 및 가공하기, 3. 프롬프트 가이드라인 제공 및 답변 수정 기능 제안)을 제안하였다.

결론 개선 전략 2의 경우 TAM(Technology Acceptance Model)의 평가 척도인 PU(지각된 유용성)에서 유의함을 보였지만, PEU(지각된 사용 용이성)의 경우 유의함을 보이지 않았는데, 이는 테스트 진행 시 타깃 유저 표집의 한계와 개선 자극물의 구동 적극성이 떨어짐이 원인이 되었음을 사후 인터뷰를 통해 확인하였다. 반면 개선 전략 1과 3의 경우 PU와 PEU 모두에서 통계적 유의성을 보임과 동시에 사후 인터뷰를 통해 긍정적 피드백을 확인할 수 있었다.

주제어 AI, Large Language Model, 유저 인터페이스

이 논문은 2023년 대한민국 교육부와 한국연구재단의 공동연구지원사업의 지원을 받아 수행된 연구임(NRF-2023S1A5A2A03084950)