*2024.02.21 | Lab Seminar*

# Understanding User Experience in Large Language Model Interactions

JIAYIN WANG, Tsinghua University, China
WEIZHI MA, Tsinghua University, China
PEIJIE SUN, Tsinghua University, China
MIN ZHANG, Tsinghua University, China
JIAN-YUN NIE, University of Montreal, Canada

HUMAN
CENTERED
COMPUTING
LABORATORY

인간중심컴퓨팅 연구실
석사과정 송형우

HUMAN
CENTERED
COMPUTING
LABORATORY

# Understanding User Experience in Large Language Model Interactions

JIAYIN WANG, Tsinghua University, China
WEIZHI MA, Tsinghua University, China
PEIJIE SUN, Tsinghua University, China
MIN ZHANG, Tsinghua University, China
JIAN-YUN NIE, University of Montreal, Canada

1. 제목만 보고

   “아니, 이렇게 큰 제목을 잡는다고? 얼마나 자신있는거지?”

2. Abstract 읽고나서

   “LLM에 대한 User-centered approach를 반복적으로 강조하네.

   얼마나 잘 이야기하는지 한 번 따져보며 읽고싶다.”

1. LLM을 사용하는 User Intent를 7개로 분류
2. LLM에 대한 User Experience에 대한 11가지 Insight를 논의
3. 이를 바탕으로 6개의 Future Research Directions 제안

# User Intent Analysis

웹 검색, 제품 검색, 멀티미디어 검색, 질문-답변 및 대화형 검색 등 다양한 정보 탐색 과정에서 사용자 의도에 대한 광범위한 연구 진행
**최근에는 생성 AI와 같은 새로운 영역에서도 일부 연구가 진행되었지만, 이는 주로 텍스트-이미지 생성 시스템이나 Bing Chat과 같은 큰 언어 모델을 활용한 검색-가이드 제품에 국한**

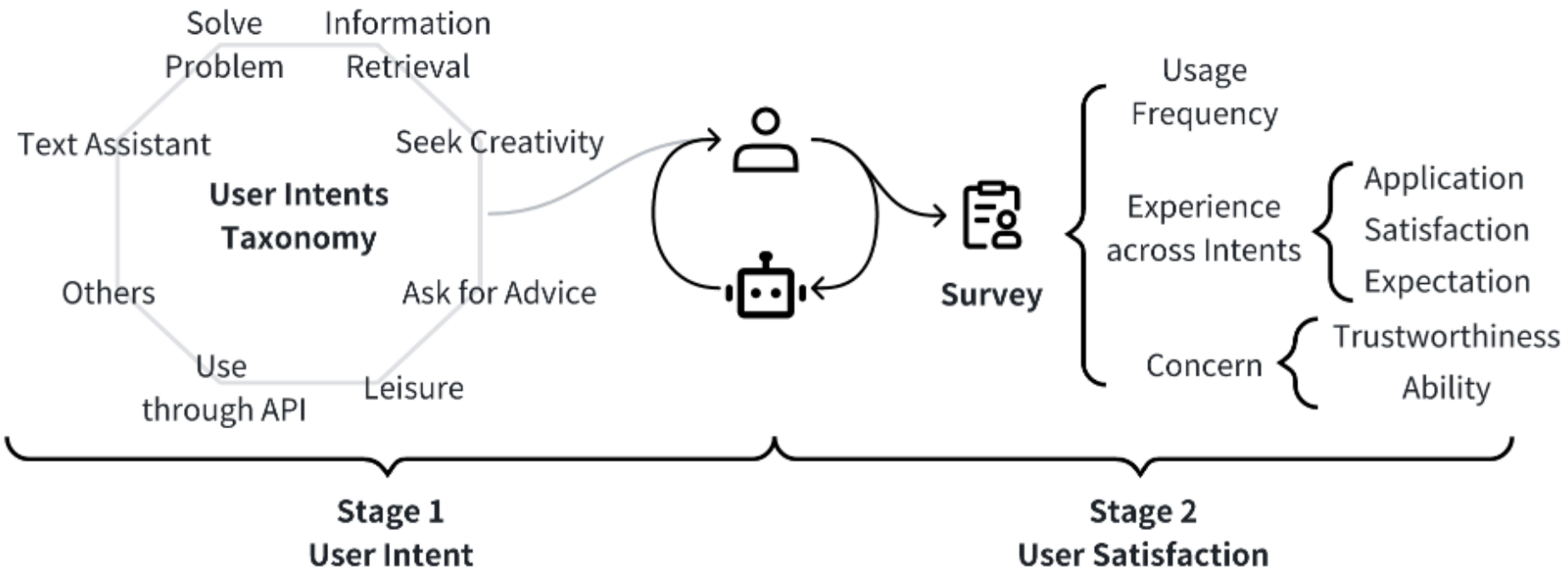
**기존 연구들은 주로 정보 탐색 행동에 초점을 맞춘 반면, 일반적인 LLM 기반 서비스에서는 더 넓은 범위의 상호작용을 다루어야 합니다.**
가장 관련성 높은 연구는 Bing Chat과 같은 검색-가이드 제품에서 나온 폐쇄 소스 로그를 사용하고 GPT-4와 인간 검증을 병행하여 의도 분류 체계 생성

## Evaluation of Large Language Models

LLM의 성능을 평가하는 수많은 benchmark들이 존재하고 있음.
**하지만, 실제로 LLM을 사용하는 사람들의 니즈를 충족하는 평가방식이 아님.**

**복잡한 real-world scenario**에서 사람들이 직접 느끼는 점을 평가해야
진정으로 LLM의 성능을 평가했다고 할 수 있음.

**LLM의 단순 성능을 높이는 것보다 User-Experience를 향상시키는 것이 중요하다**


Solve Problem
Information Retrieval
Text Assistant
Seek Creativity
User Intents Taxonomy
Others
Ask for Advice
Use through API
Leisure
Survey
Usage Frequency
Experience across Intents
Application
Satisfaction
Expectation
Concern
Trustworthiness
Ability
Stage 1 User Intent
Stage 2 User Satisfaction

## 1. User Intent 확정

**Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies**

Chirag Shah[1†‡], Ryen W. White[2†], Reid Andersen[2], Georg Buscher[2], Scott Counts[2], Sarkar Snigdha Sarathi Das[3‡], Ali Montazer[4‡], Sathish Manivannan[2], Jennifer Neville[2], Xiaochuan Ni[2], Nagu Rangan[2], Tara Safavi[2], Siddharth Suri[2], Mengting Wan[2], Leijie Wang[1‡], Longqi Yang[2]

[1]University of Washington, [2]Microsoft, [3]Pennsylvania State University, [4]University of Massachusetts Amherst
†Corresponding authors: chirags@uw.edu, ryenw@microsoft.com
‡Work done while working at Microsoft, USA

- 작년 11월에 ArXiv에 올라온 논문
- Bing 엔진의 검색 및 채팅 로그에서 사용자 의도를 분석하여 분류한 연구
- 이 때, LLM을 사용했다는 contribution이 존재함

## 1. User Intent 확정

**Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies**

Chirag Shah[1†‡], Ryen W. White[2†], Reid Andersen[2], Georg Buscher[2], Scott Counts[2], Sarkar Snigdha Sarathi Das[3‡], Ali Montazer[4‡], Sathish Manivannan[2], Jennifer Neville[2], Xiaochuan Ni[2], Nagu Rangan[2], Tara Safavi[2], Siddharth Suri[2], Mengting Wan[2], Leijie Wang[1‡], Longqi Yang[2]

[1]University of Washington, [2]Microsoft, [3]Pennsylvania State University, [4]University of Massachusetts Amherst
†Corresponding authors: chirags@uw.edu, ryenw@microsoft.com
‡Work done while working at Microsoft, USA

- Bing이기에 Information Retrieval 측면에서의 사용 용도가 많았을 것으로 예상
- General한 LLM 사용 환경에서의 Intent 분석이 필요
- ShareGPT (Open-source ChatGPR conversation logs) 활용
- 50개의 채팅 로그를 3명이서 분석하는 방법 채택

# 2. 441명 대상 User Survey

Q 1-2 **Usage Patterns**
- 1 *Services Used*
- 2 *Frequency of Usage*

Q 3-9 **User Experience across Intents**
- 3-4 *Intents Distribution*
  Choose the intents that they have used before.
  Opinions about the above intent taxonomy (optional).
- 5 *User Satisfaction across Intents*
- 6-8 *User Expectation for Different Answer Types across Intents*
  Choose between 3 pairs of answer types: detailed or concise, factual or creative, professional knowledge or common sense.
- 9 *User Expectation for Tool Utilization across Intents*
  Tools include web browsing, input analysis, personalization, programming, mathematical operations, documentation generation, and multimedia creation.

Q 10 **Anchor Question**
- 10 If the user does not follow the instructions (select B for this question), this questionnaire would be an invalid response. This helps to control the feedback quality.

Q 11 **Major Concerns**
- 11 Identify aspects of the system that need optimization, such as hallucinations, long context processing, multi-modal understanding, personalization, privacy, and safety, etc.

Q 12 **Other Comments**
- 12 Comments about the questionnaire or large language model interfaces (optional).

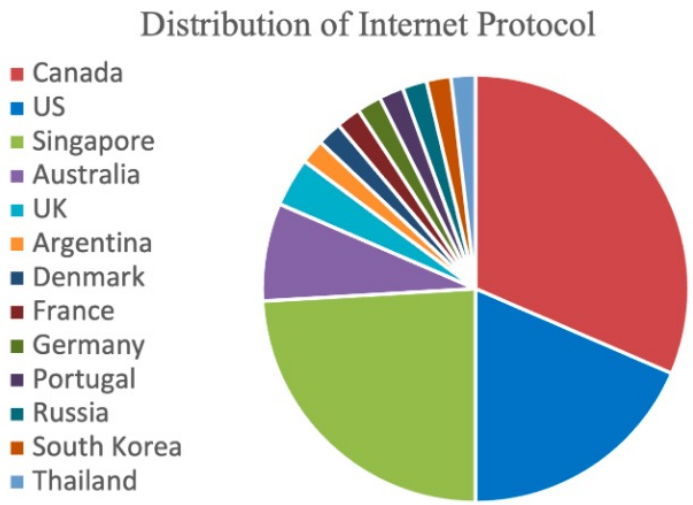
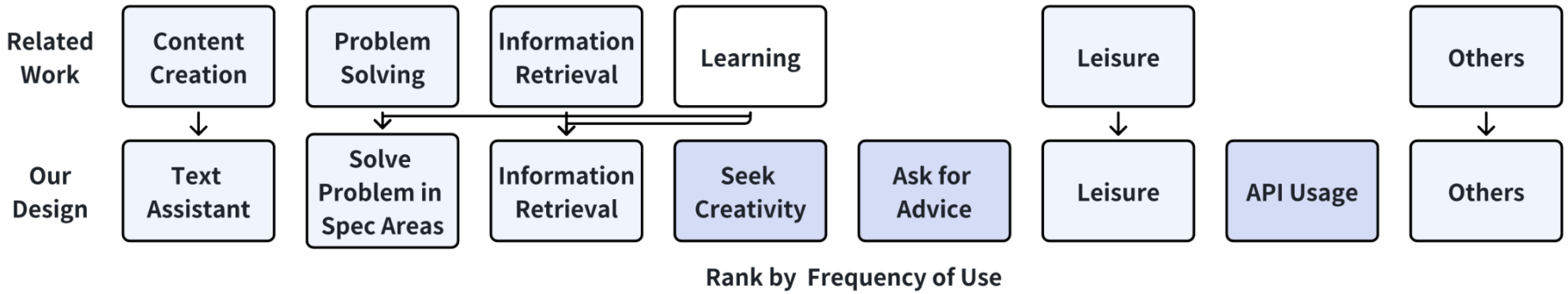
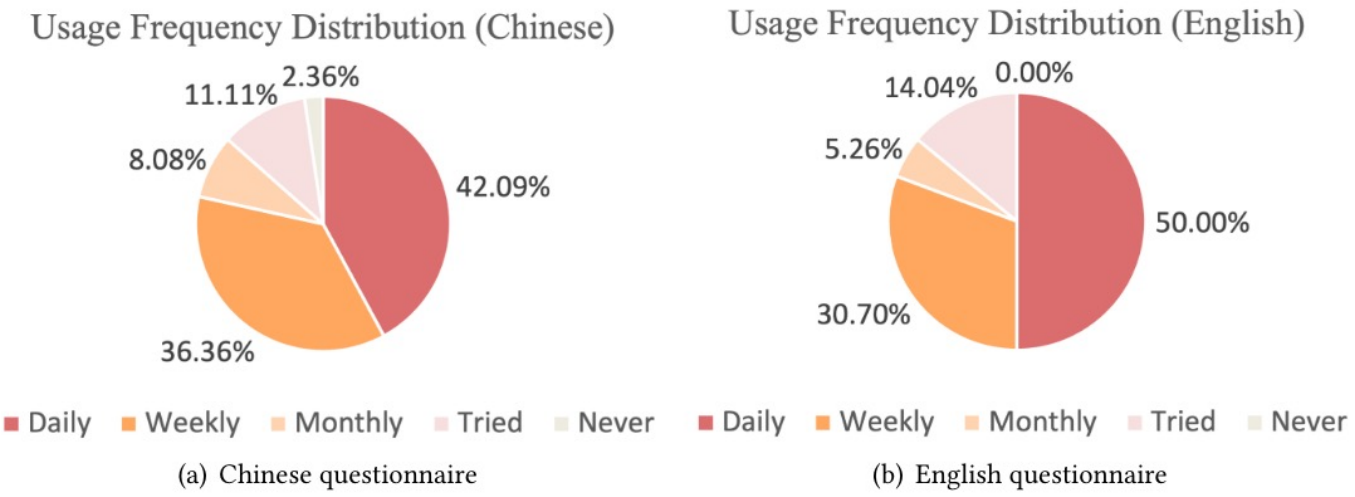



Fig. 11. IP distribution excluding China.

# 7개의 User Intent 확정



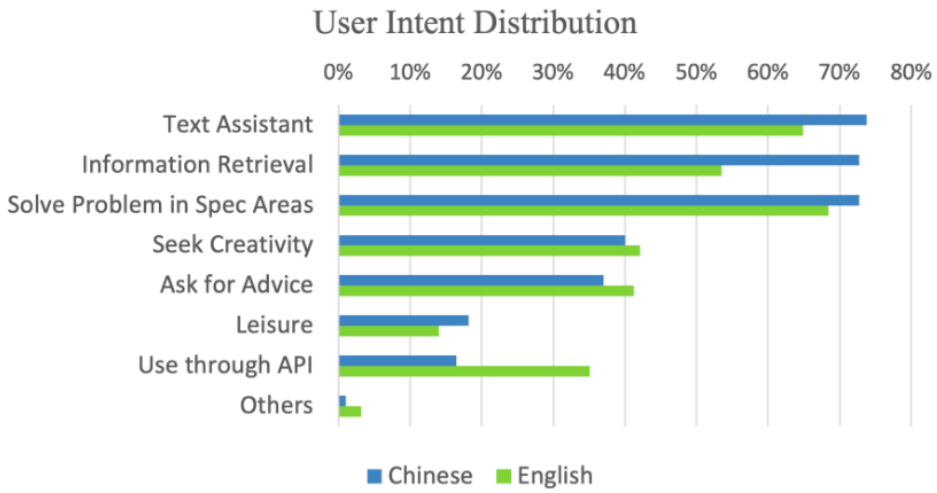
# Usage Frequency (Q2)

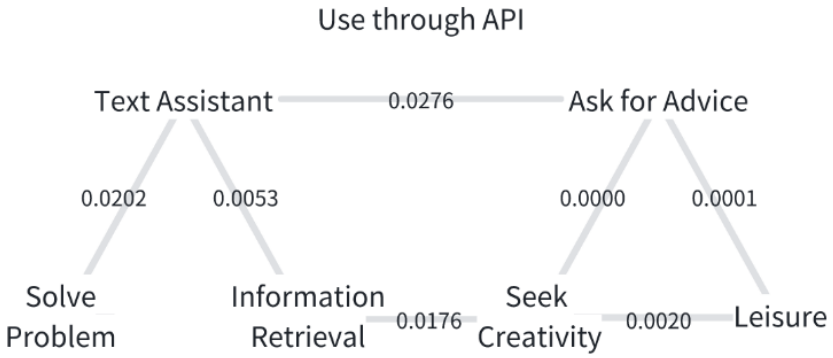



(a) Chinese questionnaire (b) English questionnaire

Finding 1: Large language model interfaces are used at least weekly by around 80% of participants.

# Intent Analysis (Q3,4)



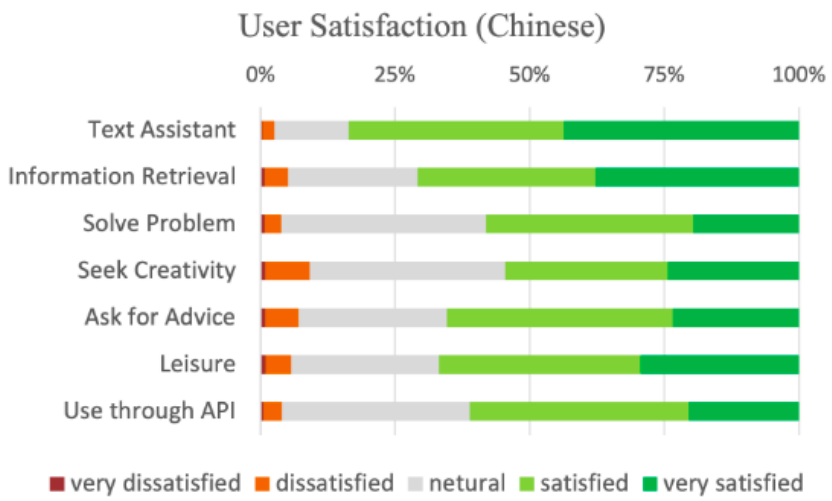
## Intent Analysis (Q3,4)

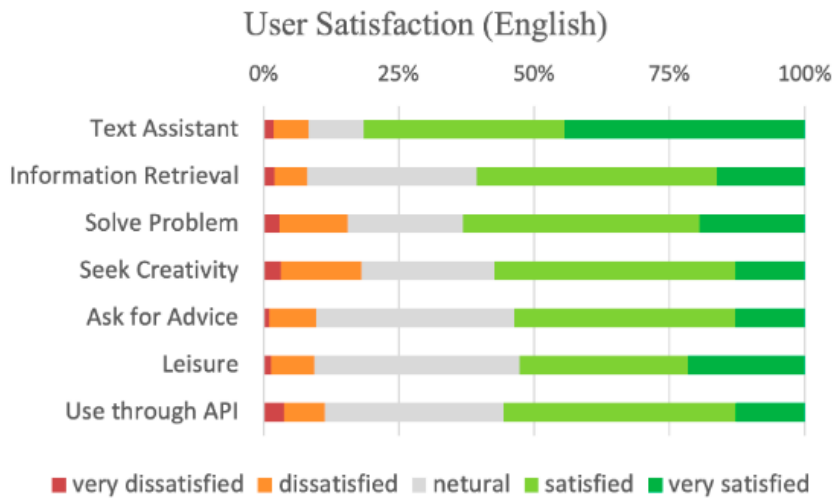
Finding 2: Based on statistical relevance, 7 intents are further clustered into 3 categories: Objective Usage via GUIs, Subjective Usage via GUIs, and Usage through APIs.

Finding 3: Text Assistant, Information Retrieval, and Solve Problem in Specialized Areas are the top three usage scenarios.

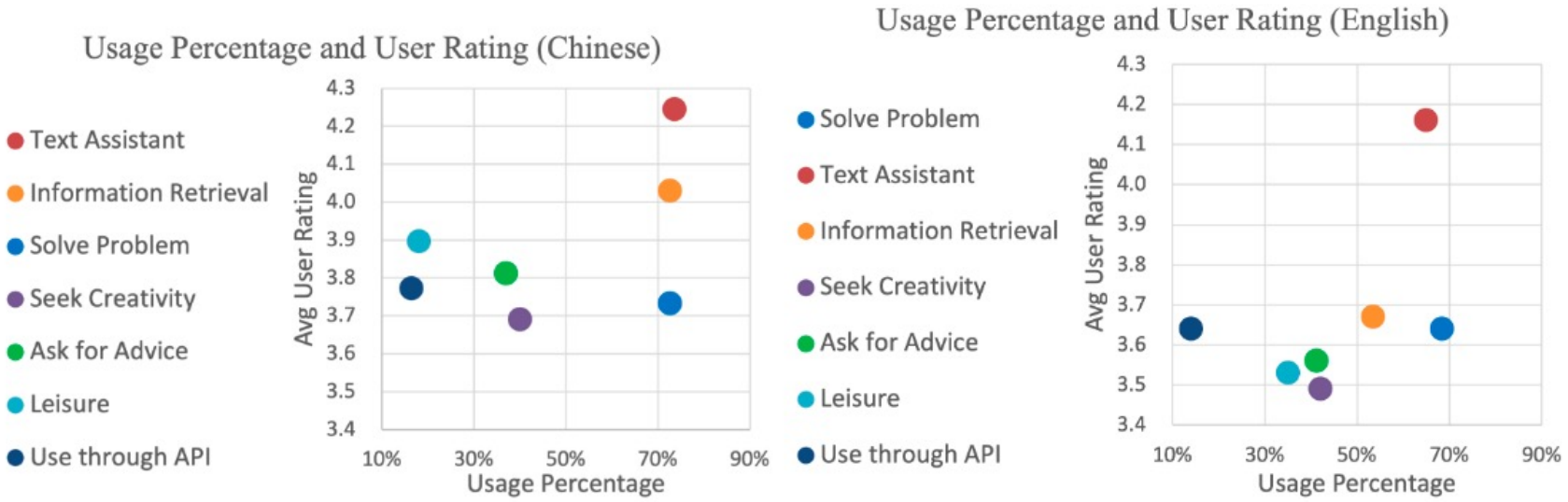
Finding 4: Subjective uses, such as Seeking Creativity and Asking for Advice, are also common intents but may have been overlooked by previous research.

# User Satisfaction (Q5)



# User Satisfaction (Q5)



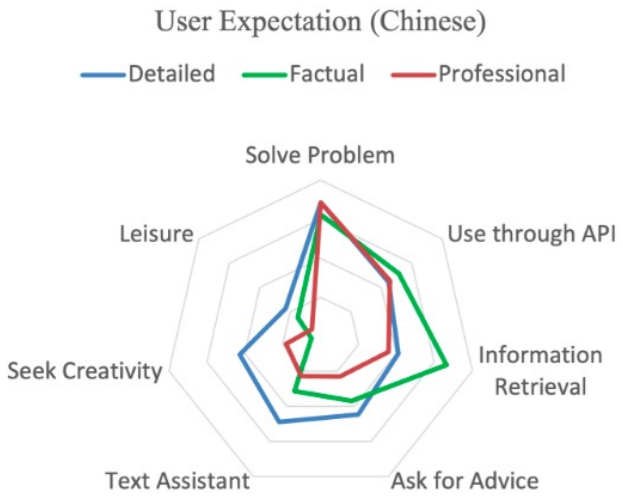
## User Satisfaction (Q5)

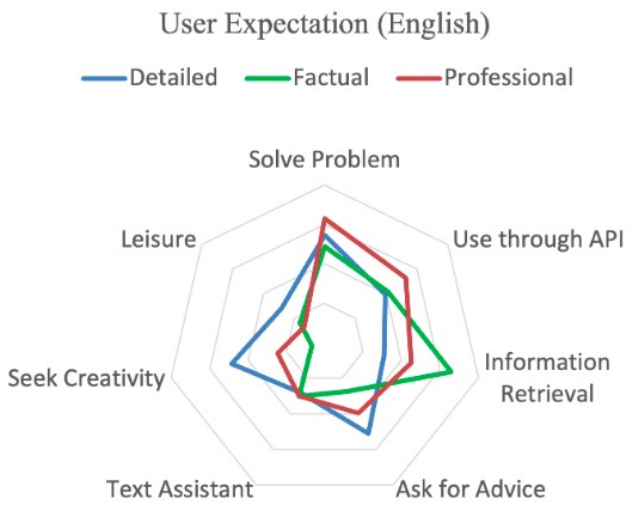
Finding 5: User studies verify that LLMs are highly effective in text manipulation tasks.

Finding 6: Subjective areas, such as Seeking Creativity, require further advances to boost user satisfaction.

Finding 7: When both frequencies of use and satisfaction are considered, they approximate a U-shape: both highly and infrequently used scenarios yield higher satisfaction levels.

# Expected Response Types (Q6,7,8)




Finding 8: User Expectations vary greatly across scenarios, which might not always align with the current evaluation standards.

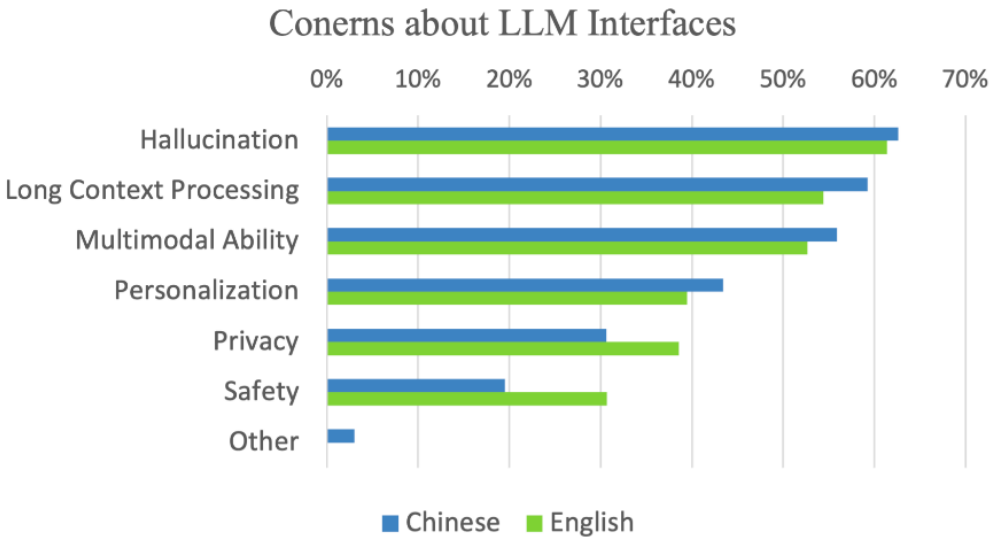
## Tool Utilization (Q9)

| | Web Browsing | Input Analysis | Personalization | Doc/PPT Generation | Multi-media | Programming | Math Operations |
|---|---|---|---|---|---|---|---|
| Text Assistant | 41% | 51% | 33% | 66% | 34% | 19% | 15% |
| Information Retrieval | 67% | 48% | 37% | 33% | 23% | 18% | 18% |
| Solve Problem | 39% | 46% | 22% | 41% | 21% | 53% | 57% |
| Seek Creativity | 47% | 42% | 55% | 40% | 53% | 18% | 17% |
| Ask for Advice | 49% | 50% | 56% | 39% | 29% | 18% | 18% |
| Leisure | 50% | 36% | 56% | 25% | 53% | 14% | 11% |
| Use through API | 30% | 43% | 24% | 34% | 23% | 42% | 38% |

Finding 9: Users anticipate specific tool utilization based on intent, underscoring the necessity of fine-grained scenario segmentation based on user intent.

Finding 10: Personalization ability is valued across all subjective usage of LLMs (Seek Creativity, Ask for Advice, and Leisure).

# Major Concerns (Q11,12)




| User Concerns in the **Others** option | Type |
|---|---|
| Professionalism, although it can answer my question, but it talks nonsense if I ask for more professional details. So this generic big model is too generic | Professionalism |
| Basically unhelpful for specialized fields | Professionalism |
| The dialog is inaccurate, the answers are too generalized, and search engines can find them as well | Inaccurate, Professionalism |
| Inaccurate | Inaccurate |
| Logic ability | Ability (logic) |
| Does not seem to recognize typos very well in Chinese | Ability (linguistic) |
| Low degree of freedom and many restrictions | Freedom |

Finding 11: The user concerns and desired improvements are mainly two parts: model capability and trustworthiness.

1. **사용자 중심의 평가 프레임워크의 필요성**
    1. 실제 애플리케이션과 더 관련성이 높은 테스트 시나리오와 지표를 사용해야 함
2. **AGI로 나아가기 위해 LLM 사용 의도에 대해 더 분석해야할 필요성**
    1. LLM을 다양한 시나리오에 맞춰 더욱 개선할 수 있고 다양하고 광범위한 데이터를 학습할 수 있는 기회 생성 가능함

1. LLM을 사용하는 User Intent를 7개로 분류
2. LLM에 대한 User Experience에 대한 11가지 Insight를 논의
3. 이를 바탕으로 6개의 Future Research Directions 제안

행복한 하루 되세요