# **시장 조사에서 대규모 언어 모델(LLM)의 활용 및 소비자 선호도 분석에 대한 보고서**

## **핵심 요약**

본 보고서는 대규모 언어 모델(LLM), 특히 GPT-3.5 Turbo가 시장 조사 도구로서 소비자 선호도 분석 및 지불 의사(WTP) 추정에 어떻게 활용될 수 있는지에 대한 연구를 심층적으로 분석한다. 이 연구는 GPT-3.5 Turbo에 설문조사 질문을 질의하여 다수의 응답을 생성하고, 이를 인간 설문조사 데이터와 비교하는 실증적 접근 방식을 채택했다. LLM 응답의 분포적 특성과 미세 조정(fine-tuning)이 미치는 영향에 중점을 두었다.

주요 발견 사항은 두 가지로 요약될 수 있다. 첫째, GPT가 생성한 제품 및 기능에 대한 WTP 추정치는 종종 현실적이며, 특정 미세 조정 없이도 인간 연구의 추정치와 유사한 규모를 보였다.1 그러나 완전히 새로운 기능, 특정 브랜드 인식(예: MacBook에 대한 부정적인 WTP), 또는 직관에 반하는 속성 가치 평가(예: 더 작은 노트북 화면 선호)에서는 상당한 차이가 발생할 수 있다.1 둘째, 이전 설문조사 데이터를 미세 조정에 활용하면 GPT의 응답이 크게 향상되며, 특히

*동일한 제품 범주 내의 새로운 기능*에 대한 인간 선호도와의 일치도가 현저히 개선되었다.1 이 방법은 GPT의 초기 오판(예: 새로운 맛에 대한 WTP 부호 변경)까지도 수정할 수 있음을 보여주었다.1

이러한 발전에도 불구하고, LLM은 현재 *완전히 새로운 제품 범주*로 학습된 선호도를 효과적으로 확장하는 데 어려움을 겪으며(이 경우 미세 조정이 오히려 일치도를 악화시킬 수 있음), 다양한 고객 세그먼트(예: 소득 또는 성별)에 걸쳐 *이질적인 선호도를 정확히 포착*하는 데 한계가 있다.1 LLM은 주로 평균적인 인구의 WTP를 반영하는 경향이 있다.1

이러한 결과는 LLM이 시장 조사자에게 빠르고 저렴한 통찰력을 제공하는 귀중한 *보완적* 도구임을 시사한다.1 LLM의 최적의 활용은 특히 이전 설문조사 데이터가 있는 경우, 제품 개발 초기 단계에서 새로운 기능 아이디어를 테스트하고 개선하는 데 있다.1 그러나 LLM은 미묘한 세그먼트 차이를 이해하거나 완전히 새로운 제품 범주를 평가하는 데 있어 인간 응답자를 완전히 대체할 수는 없다.1

LLM이 "프록시 소비자" 역할을 수행할 수 있다는 점은 중요하게 다루어진다. 연구는 GPT가 "현실적인" WTP 추정치를 생성하고 "학습된 소비자 응답을 반영"한다고 강조한다.1 이는 GPT가 일반적인 소비자의 대리 역할을 할 수 있음을 의미한다. 그러나 새로운 범주와 이질성에 대한 어려움은 이러한 대리 역할이 제한적임을 보여준다. 즉, GPT는 광범위하고 집계된 선호도를 반영하는 데는 능숙하지만, 진정으로 새로운 통찰력이나 세그먼트별 깊이 있는 이해를 제공하는 데 필요한 인간 경험의 깊이는 부족하다. 따라서 시장 조사자들은 GPT를 개별 소비자의 목소리 대체물이라기보다는, 기존 공개 데이터를 신속하게 종합하여 그럴듯한 WTP 추정치를 도출하는 정교한 "시장 정보 집계자"로 보아야 한다. 그 유용성은 광범위한 경향 파악과 반복적인 테스트에 있으며, 깊이 있는 질적 이해나 정밀한 세분화에는 적합하지 않다.

이러한 분석은 LLM의 유용성이 발휘되는 "골디락스 존"이 존재함을 보여준다. LLM은 초기에는 합리적인 성능을 보이며, *동일 범주 내의 새로운 기능*에 대한 미세 조정을 통해 크게 개선되지만, *새로운 범주*와 *이질성*에는 실패한다.1 이는 LLM이 가장 효과적인 영역을 명확히 정의한다: 기존의 관련 데이터를 미세 조정에 활용할 수 있고, 알려진 영역 내에서 점진적인 혁신을 다루는 작업이다. 이 영역을 벗어나면 LLM의 유용성은 급격히 감소한다. 이는 단순히 한계가 아니라, LLM의 최적 적용 공간을 규정하는 요소이다. 이러한 "골디락스 존"은 시장 조사에서 LLM의 전략적 배포 모델을 제안한다: 기존 지식 기반을 활용하여 기능 확장의 신속한 프로토타이핑 및 검증에 LLM을 사용해야 한다. 반면, 인간의 상당한 개입 없이 완전히 새로운 아이디어를 탐색하거나 틈새시장을 깊이 파고드는 데는 사용을 지양해야 한다. 이는 "보완 대 대체" 논쟁을 "정확히 어디를 보완할 것인가"라는 보다 미묘한 전략으로 재구성한다.

## **서론: 시장 조사에서 LLM 활용**

대규모 언어 모델(LLM)은 방대한 양의 텍스트 데이터를 학습하여 인간과 유사한 언어를 이해하고 생성하도록 설계된 인공지능의 한 유형이다.1 이러한 모델은 언어 번역, 음성 인식, 콘텐츠 생성, 텍스트 분류 등 다양한 산업 분야에서 광범위하게 활용되며 그 인기가 빠르게 증가하고 있다.1 본 연구는 LLM, 특히 GPT가 시장 조사 도구로서의 잠재력을 탐구한다.

이 연구의 핵심 가설은 GPT가 시장 조사 설문조사에 응답할 때, 제품 리뷰나 온라인 포럼과 같은 방대한 훈련 데이터에서 학습된 소비자 응답을 반영할 것이라는 점이다.1 이는 GPT가 소비자 선호도에 대한 귀중하고 빠르며 저렴한 통찰력을 제공하여, 컨조인트 분석이나 포커스 그룹과 같은 비용이 많이 드는 전통적인 방법을 보완하거나 잠재적으로 대체할 수 있음을 시사한다.1 그러나 연구는 GPT 응답의 품질과 일관성, 특히 가격 책정 및 질문 간의 내부 일관성에 대한 불확실성을 인정한다.1 또한 설문조사에서 얻은 "진술된 선호도"와 실제 수요 데이터에서 나타나는 "드러난 선호도" 간의 차이점을 논의하며, GPT의 훈련 세트가 이 두 가지 측면을 모두 포함하고 있음을 언급한다.1

GPT의 훈련 세트가 "진술된 선호도"(설문조사, 리뷰)와 "드러난 선호도"(실제 수요 데이터, 온라인 행동을 통해 암시됨)를 모두 포함한다는 점은 중요한 고려 사항이다.1 경제학과 마케팅에서 이 둘은 중요한 차이를 가진다. LLM이 이 둘의 혼합으로 훈련되었다면, 그 결과물은 잠재적으로 선택의 근본적인 동인을 모호하게 만들 수 있는 혼합된 형태일 수 있다. 진술된 선호도는 이상적이거나 편향될 수 있는 반면, 드러난 선호도는 실제 행동을 반영한다. GPT가 인기 브랜드(MacBook, iPad)에 대해 때때로 부정적인 WTP를 제안하는 것은 훈련 데이터의 "드러난 선호도" 편향(예: 널리 사용되는 제품에 대한 더 많은 부정적인 리뷰, 또는 실제 구매 데이터에서 가격 민감도 반영)에서 비롯될 수 있다. 따라서 연구자들은 LLM이 생성한 선호도가 순수한

*진술된 의도*를 반영하는 것이 아니라, 온라인 담론과 추론된 행동의 복합적인 집합체일 수 있음을 인지해야 한다. 이는 LLM이 단순히 "설문조사 응답자를 시뮬레이션"하는 것이 아니라, 잠재적으로 "집계된 온라인 시장 정서를 시뮬레이션"하는 것이며, 이는 비록 가치 있을지라도 다른 구성체가 될 수 있음을 의미한다.

이 연구의 실증 분석은 GPT가 다양한 수준의 사전 정보를 가지고 소비자 선호도를 시뮬레이션하는 능력과 인간 설문조사 데이터가 GPT를 어떻게 보완할 수 있는지를 탐구함으로써 실제 시장 조사 문제를 근사화하는 것을 목표로 한다.1 전반적인 결과는 GPT에서 도출된 WTP 추정치가 현실적인 규모와 분포를 가지며 인간 설문조사와 비교할 만하다는 것을 보여준다.1 인간 설문조사 데이터로 GPT를 보강하면 해당 설문조사와의 일관성이 향상되고 새로운 제품 기능에 대한 인간 선호도와 더 잘 일치한다.1 그러나 이 접근 방식은 다른 제품 범주에 대해서는 더 나은 일치도를 보이지 않는다.1 연구는 GPT가 인간 테스트 전에 새로운 기능 아이디어를 테스트하고 좁히는 데 도움이 될 수 있다고 제안하며, 특히 이전 설문조사 데이터가 있을 때 유용하다고 강조한다.1 또한 GPT는 현재 인구 통계학적 그룹 간의 이질적인 선호도를 의미 있게 반영하는 데 어려움을 겪으며, 주로 평균 인구 WTP를 반영한다고 지적한다.1

연구는 GPT 응답이 "훈련 데이터에서 소비자 응답을 반영할 것"이라는 가설을 세운다.1 이는 근본적인 전제이다. 그러나 LLM은 추론하는 존재가 아니라 패턴 매칭 기계이다. 인간의 추론을

*모방*할 수는 있지만, 인간적인 의미에서 "가치"나 "선호도"와 같은 개념을 *이해*하지는 못한다. 연구 설계에서 관찰된 "더 작은 표본 크기에서의 일관성"(GPT 응답을 300개에서 50개로 줄임)은 GPT의 "시뮬레이션된 선택" 내에서 높은 수준의 내부 일관성을 의미하며, 이는 인간 의사 결정의 자연스러운 변동성과 때때로 나타나는 비합리성을 완전히 반영하지 못할 수 있다. 따라서 시장 조사에서 LLM의 유용성은 인간 응답을 설득력 있게 *시뮬레이션*하는 능력에 달려 있으며, 반드시 인간의 사고 과정을 *복제*하는 것은 아니다. 이는 결과물이 그럴듯해 보일지라도, 그 기저의 "논리"는 통계적 패턴 인식을 기반으로 한다는 것을 의미한다. 이는 특히 LLM 결과물이 인간의 직관이나 확립된 시장 이해와 크게 다를 때(예: 인기 브랜드에 대한 부정적인 WTP) 인간의 검증과 해석이 필요하다는 점을 강조한다. 또한 LLM은 방대한 텍스트 데이터를 기반으로 "평균적인 사람이 *무엇을 말할 것인가*를 예측"하는 데 더 능숙하며, "왜 그렇게 말할 것인가를 이해"하는 데는 부족하다.

이 연구의 핵심 방법론적 접근 방식은 GPT를 인간 응답의 시뮬레이터로 사용하는 데 중점을 둔다.1 컨조인트 분석은 인간 및 GPT 생성 응답 모두에서 제품 속성에 대한 WTP를 추정하는 데 사용된다.1 또한 인구 통계학적 이질성을 시뮬레이션하고 기존 설문조사 데이터를 통한 미세 조정을 통합하는 방법을 체계적으로 탐구한다.1

## **주요 연구 결과: LLM 기반 소비자 선호도 분석**

### **GPT의 WTP 추정치와 인간 데이터 비교**

LLM을 활용한 컨조인트 분석의 유용성을 평가하기 위해, GPT가 생성한 WTP 추정치를 인간 설문조사 결과와 비교하고, 소비자 이질성을 반영하는 GPT의 성능을 검토하며, 미세 조정의 영향을 분석했다.

#### **초기 성능 (기준선)**

초기 파일럿 연구에서는 이전 모델인 text-davinci-003을 사용하여 치약과 데오도란트 같은 가정용 소비재를 분석했다. 불소와 알루미늄에 대한 WTP 추정치는 Fong et al. (2023)의 인간 기준선과 비교할 때 50% 이내로 근접하여 LLM의 시장 조사 유용성에 대한 유망한 시작점을 보여주었다.1

주요 연구에서는 GPT-3.5 Turbo를 사용하여 치약(연구 1A)에 대한 GPT의 WTP 추정치가 거의 모든 속성의 순위를 정확하게 매겼다.1 시나몬 및 딸기 맛과 같은 속성에 대한 추정된 WTP 규모는 인간 응답과 유사했지만, 불소에 대한 GPT의 WTP는 인간보다 현저히 높게 나타났다.1 이러한 초기 결과는 LLM이 특정 미세 조정 없이도 제품 속성과 그 일반적인 인지 가치에 대한 기본적인 이해를 가지고 있음을 시사한다. 이는 LLM이 방대한 훈련 데이터에서 얻은 광범위한 공개 담론을 반영한 것으로 보이며, 시장 선호도에 대한 합리적인 첫 번째 근사치를 제공할 수 있음을 의미한다.

#### **상당한 불일치 및 미묘한 차이**

새로운 치약 맛(연구 1B)의 경우, 오이 및 팬케이크 맛에 대한 GPT의 WTP 평가는 인간 데이터와 특히 WTP의 *부호*에서 크게 달랐다(GPT는 종종 긍정적, 인간은 부정적).1 이러한 불일치는 LLM 내의 "신제품 시도" 편향 또는 진정으로 새로운 개념에 대한 부정적인 연관성을 형성할 충분한 실제 데이터 부족에서 기인할 수 있다. 인간이 맛에 대한 즉각적인 혐오감을 가질 수 있는 것과 달리, GPT의 패턴 매칭은 "새로운 것"을 "잠재적으로 흥미로운 것"으로 해석하거나 단순히 강한 부정적 선호도를 형성할 만큼 충분한 부정적 데이터 포인트가 부족했을 수 있다. 따라서 LLM을 완전히 새로운 개념의 아이디어 구상이나 초기 단계 테스트에 사용할 때, 시장 조사자들은 인위적으로 부풀려진 긍정적인 반응에 주의해야 한다. 매우 혁신적이거나 비전통적인 기능에 대한 LLM 결과는 추가적인 정밀 검토와 초기 인간 검증이 필요할 수 있는데, 이는 LLM이 진정으로 낯선 것에 대한 인간의 혐오감이나 회의론을 정확하게 포착하지 못할 수 있기 때문이다.

소비자 전자제품 연구(노트북 및 태블릿)에서는 두드러진 차이가 관찰되었다. GPT는 MacBook 브랜드(연구 2A, 2B)에 대해 부정적인 WTP를, iPad(연구 2C)에 대해서는 인간 선호도에 비해 상대적으로 낮은 선호도를 암시했다.1 이는 일반적인 시장 인식 및 확립된 브랜드 가치와 직접적으로 상충된다. GPT의 MacBook에 대한 부정적인 WTP와 iPad에 대한 낮은 선호도는 강력한 시장 지위와 브랜드 가치에도 불구하고 상당한 차이를 보인다. 이는 여러 요인에서 비롯될 수 있다. 첫째, 훈련 데이터 편향: 온라인 토론(훈련 데이터의 상당 부분)은 Apple과 같이 널리 사용되고 가시성이 높은 제품에 대해 더 많은 불만이나 비판적인 리뷰를 포함할 수 있다. 이러한 "노이즈가 많은" 데이터는 LLM의 인식을 왜곡할 수 있다. 둘째, 경험적 이해 부족: LLM은 브랜드 충성도, 디자인 미학 또는 사용자 생태계의 이점을

*경험*하지 않는다. 그들의 "이해"는 순전히 텍스트 기반이다. 그들은 브랜드 가치의 무형적 가치보다 가격 민감도에 대한 논의를 더 많이 포착할 수 있다. 셋째, 컨조인트 설계 한계(LLM의 경우): 컨조인트는 인간에게는 좋지만, LLM은 단순화된 속성 세트에서 브랜드의 복잡하고 전체적인 가치를 추론하는 데 어려움을 겪을 수 있다. 따라서 LLM은 현재 브랜드 가치나 소비자가 제품에 대해 가지는 깊은 감정적 연결을 평가하는 데는 부적합하다. 시장 조사자들은 브랜드 가치 또는 브랜드별 기능 평가 시 LLM을 신중하게 사용해야 한다. 인간 연구는 브랜드 인식의 질적 및 감정적 측면을 이해하는 데 여전히 중요하다. 이는 또한 LLM이 *기능적 속성 간의 상충 관계*를 처리하는 데 더 능숙하며, *전체적인 브랜드 인식*에는 부족함을 시사한다.

노트북의 "내장 프로젝터"(연구 2B)에 대한 GPT의 WTP는 인간 추정치의 거의 3배에 달했다.1 또한 GPT는 더 큰 노트북 화면 크기에 대해 부정적인 WTP를 제안하며 13인치 화면을 선호했는데, 이는 일반적인 소비자 트렌드와도 다르다.1 이러한 불일치는 LLM이 사전 훈련된 지식에만 의존할 때의 한계를 부각시킨다. 이는 훈련 데이터에 존재하는 편향(예: 광범위한 사용으로 인해 매우 인기 있는 브랜드에 대한 더 비판적인 온라인 논의, 또는 텍스트 설명 이상의 기능 유용성에 대한 미묘한 이해 부족)을 반영할 수 있다. 전반적으로, GPT-3.5 Turbo는 많은 속성에 대해 유사한 규모와 부호의 WTP 추정치를 제공하지만, 특히 새로운 기능에 대해서는 상당한 차이가 존재하여, LLM이 유용한 신호는 될 수 있으나 인간 응답을 완전히 대체할 수는 없음을 시사한다.1

#### **LLM 선택에 대한 견고성**

연구는 노트북 연구(연구 2A)를 GPT-4o, Claude 3 Haiku, Claude 3.5 Sonnet, LLaMA-3 (8B 및 70B)를 포함한 다양한 LLM으로 재현했다.1 결과는 GPT-3.5 Turbo와 GPT-4o가 Claude 및 LLaMA-8B보다 인간 응답에 더 가깝게 근사했음을 나타냈다.1 GPT-4o와 GPT-3.5 Turbo 사이에도 차이가 존재했지만, 미세 조정이 이러한 차이를 줄일 것이라는 가설이 제시되었다.1

다음 표는 GPT-3.5 Turbo와 인간 설문조사 간의 선택된 제품 속성에 대한 WTP 추정치를 비교한다.

**표 1: 선택된 제품 속성에 대한 WTP 추정치 비교: GPT-3.5 Turbo vs. 인간 설문조사**

| 제품 범주/연구 | 속성 | 인간 WTP 추정치 (USD) | GPT-3.5 Turbo WTP 추정치 (USD) | 질적 평가 |
| --- | --- | --- | --- | --- |
| 치약 연구 1A | 불소 | - | 높음 | GPT WTP가 인간보다 높음 |
| 치약 연구 1A | 시나몬 맛 | 유사 | 유사 | 유사한 규모 |
| 치약 연구 1A | 딸기 맛 | 유사 | 유사 | 유사한 규모 |
| 치약 연구 1B | 오이 맛 | 부정적 | 긍정적 | WTP 부호 다름 (GPT는 긍정적) |
| 치약 연구 1B | 팬케이크 맛 | 부정적 | 긍정적 | WTP 부호 다름 (GPT는 긍정적) |
| 노트북 연구 2A, 2B | MacBook 브랜드 | 긍정적 | 부정적 | GPT는 부정적인 WTP 암시 |
| 노트북 연구 2C | iPad 브랜드 | 긍정적 | 낮음 | GPT는 상대적으로 낮은 선호도 암시 |
| 노트북 연구 2B | 내장 프로젝터 | 낮음 | 거의 3배 높음 | GPT WTP가 인간보다 현저히 높음 |
| 노트북 연구 2A | 13인치 화면 | 긍정적 | 긍정적 | GPT는 더 큰 화면에 부정적 WTP 암시 |

### **미세 조정(Fine-tuning)의 효과 및 적용**

관련 인간 설문조사로 GPT를 미세 조정하는 것이 일치도를 개선하는 데 미치는 영향을 탐구했다.

#### **동일 범주 내의 새로운 기능에 대한 일치도 개선**

연구 결과, 관련 인간 설문조사 데이터로 GPT를 미세 조정하면 WTP 추정치가 크게 향상되는 것으로 나타났다.1 예를 들어, 연구 1A 데이터(기존 치약)로 GPT를 미세 조정하면 연구 1B(새로운 치약 맛)에서 불소, 시나몬, 딸기 맛에 대한 WTP가 인간 수준에 더 가깝게 개선되었다.1 특히, 이 미세 조정은 오이 및 팬케이크 맛에 대한 WTP

*부호를 긍정에서 부정으로 변경*하여 인간 선호도와 일치시켰다.1 마찬가지로, 연구 2A 데이터(노트북)로 연구 2B(새로운 기능이 있는 노트북)를 미세 조정하면 "내장 프로젝터"에 대한 일치도가 크게 개선되어 GPT의 WTP 추정치가 인간과 매우 유사해졌다.1

이는 특정 인간 데이터가 LLM의 방대한 일반 지식을 특정 도메인에 "고정"하거나 "안내"하여 점진적인 혁신에 대한 예측을 더 정확하게 만드는 강력한 시연이다. 이는 LLM이 관련성 있고 고품질의 도메인 내 데이터를 제공받을 때 특정 시장 맥락에서 효과적으로 학습하고 초기 편향을 극복할 수 있음을 시사한다. 미세 조정의 성공은 동일 범주 내의 새로운 기능에 대한 미세 조정이 성공한 반면, 새로운 범주에 대해서는 실패했다는 점에서 "맥락적 기반 다지기" 메커니즘으로 작용한다는 것을 보여준다. 미세 조정은 LLM에 완전히 새로운 개념을 가르치는 것이 아니라, 기존의 방대한 지식을 특정하고 좁은 도메인에 맞게 재조정하는 데 도움을 준다. 도메인이 너무 극적으로 변하면(새로운 제품 범주), 기존의 "기반 다지기"는 무관해지거나 심지어 역효과를 내어 부정적인 전이 학습으로 이어진다. 따라서 시장 조사자들은 미세 조정을 모든 LLM 한계에 대한 만병통치약으로 보는 대신, 잘 정의된 제품 공간 내에서 정확성과 관련성을 높이는 방법으로 보아야 한다. 이는 기존 지식을 정제하는 것이지, 처음부터 새로운 도메인 전문 지식을 창출하는 것이 아니다. 이는 *미세 조정 데이터*의 품질과 관련성이 가장 중요하다는 것을 의미한다.

#### **새로운 제품 범주로의 제한된 확장**

완전히 새로운 제품 범주로 확장하려는 시도에서는 중요한 한계가 관찰되었다.1 연구 2C에서 태블릿 질의를 위해 노트북 데이터로 GPT를 미세 조정했을 때, 브랜드, 저장 공간 및 RAM에 대한 인간과 GPT WTP 추정치 간의

*거리가 증가*했으며, 크기에 대한 부호까지 바뀌었다.1

이는 표면적으로 유사한 속성을 가질지라도, 서로 다른 제품 범주 간의 근본적인 맥락적 차이(예: 사용 패턴, 폼 팩터, 경쟁 환경, 소비자 기대치)가 LLM이 단순한 미세 조정을 통해 효과적으로 연결하기에는 너무 복잡하다는 것을 시사한다. 이는 관련 없는 도메인의 미세 조정 데이터가 오히려 성능을 저하시키는 "부정적인 전이 학습"으로 이어질 수 있다. 범주 간 미세 조정의 부정적인 영향은 현재 LLM의 근본적인 "도메인 특이성" 장벽을 지적한다. 속성(저장 공간, RAM)이 표면적으로 유사하더라도, 그들의 *의미와 가치*는 다른 제품 유형에 따라 크게 달라진다. 노트북의 16GB RAM은 태블릿의 그것과 사용자 기대치 및 인지 가치 측면에서 다른 의미를 가진다. LLM은 방대한 훈련에도 불구하고, 해당 *새로운 도메인 내*에서 명시적이고 방대한 훈련 없이는 이러한 미묘한 도메인별 관계를 추론하는 데 어려움을 겪는다. 이는 LLM이 강력한 일반론자이지만, 완전히 다른 제품 생태계에 걸쳐 속성 가치를 추상화하는 데는 아직 능숙하지 않다는 것을 시사한다. 시장 조사자들은 범주 간 미세 조정을 피해야 한다. 이는 LLM이 알려진 도메인 내에서 *내삽*하는 데는 강력한 도구이지만, 진정으로 새로운 도메인으로 *외삽*하는 데는 약하다는 생각을 강화한다.

#### **LLM 견고성에 미치는 영향**

연구는 GPT-4o로 프로젝터 연구를 *미세 조정 없이* 재현했을 때도 인간 및 미세 조정된 GPT-3.5 Turbo와 유사한 WTP 추정치를 생성했다는 점에 주목했다.1 이는 미세 조정이 LLM 간의 내재된 차이를 줄이고 일관성을 개선하여, 특정 도메인 지식을 제공받을 때 그 결과물을 인간 선호도에 가깝게 "정규화"할 수 있다는 가설을 뒷받침한다.1

다음 표는 미세 조정이 WTP 일치도에 미치는 영향을 새로운 기능과 새로운 카테고리에 걸쳐 보여준다.

**표 2: 미세 조정이 WTP 일치도에 미치는 영향: 신규 기능 vs. 신규 카테고리**

| 미세 조정 시나리오 | 대상 속성 | 인간 WTP (USD) | GPT WTP (미세 조정 없음) (USD) | GPT WTP (미세 조정 있음) (USD) | 결과 |
| --- | --- | --- | --- | --- | --- |
| 치약 1A FT for 1B (새로운 맛) | 오이 맛 | 부정적 | 긍정적 | 부정적 | 부호가 인간과 일치하도록 변경됨, 일치도 개선 |
| 치약 1A FT for 1B (새로운 맛) | 팬케이크 맛 | 부정적 | 긍정적 | 부정적 | 부호가 인간과 일치하도록 변경됨, 일치도 개선 |
| 노트북 2A FT for 2B (프로젝터) | 내장 프로젝터 | 낮음 | 거의 3배 높음 | 인간과 유사 | 상당한 개선 |
| 노트북 2A FT for 2C (태블릿) | MacBook 브랜드 (태블릿) | 긍정적 | 부정적 | 더 부정적 | 불일치 증가 |
| 노트북 2A FT for 2C (태블릿) | 저장 공간 (태블릿) | 긍정적 | 긍정적 | 불일치 증가 | 불일치 증가 |
| 노트북 2A FT for 2C (태블릿) | RAM (태블릿) | 긍정적 | 긍정적 | 불일치 증가 | 불일치 증가 |
| 노트북 2A FT for 2C (태블릿) | 크기 (태블릿) | 긍정적 | 부정적 | 부호 변경 | 불일치 증가 |

### **소비자 이질성 반영의 한계**

LLM의 중요한 한계는 다양한 인구 통계학적 그룹에 걸쳐 이질적인 선호도를 의미 있게 반영하는 데 어려움을 겪는다는 점이다.1 인간 설문조사 데이터에서 세그먼트별로 통계적으로 유의미한 WTP 차이가 나타났음에도 불구하고(예: 소득, 성별 또는 인종별), GPT는 이러한 차이를 정확하게 포착하지 못했다.1

연구 2A에서 RAM이나 저장 공간에 대한 인간 WTP는 소득 그룹 간에 거의 차이가 없었지만, GPT는 큰 차이와 현저히 다른 옵트아웃 비율을 반영했다.1 MacBook WTP의 경우, GPT 추정치는 인간 데이터와 규모 및 부호에서 달랐지만, 그룹 간(예: 소득별)의 일관된

*순서*를 보였다.1 그러나 GPT는 인간 데이터에서 관찰된 미묘한 패턴과 달리 이 순서를

*모든 속성*에 일반화하는 경향이 있었다.1 전반적으로, GPT는 인간 표본에서 발견된 차이점 중 일부만 반영했으며, 특정 그룹의 선호도를 잘 반영하지 못했다.1

심지어 인간 설문조사의 인구 통계학적 정보(예: 연구 2A를 연구 2B 질문에 미세 조정)로 GPT를 미세 조정했을 때도, WTP 추정치는 *평균* 인간 추정치와 더 유사해지고 소득 분포에 걸쳐 MacBook 선호도의 부호를 올바르게 변경했지만, GPT는 여전히 기존 또는 새로운 속성에 대해 통계적으로 유의미한 인간 데이터의 차이에도 불구하고 인구 통계학적 그룹 간의 선호도를 의미 있게 복구하고 차별화할 수 없었다.1 이는 시장 세분화, 타겟 마케팅 및 개인화된 제품 개발에 있어 주요 한계이다. LLM은 다양한 텍스트로 훈련되었음에도 불구하고 주로 평균 인구 선호도를 학습하는 것으로 보인다. 별개의 인구 통계학적 세그먼트를 정확하게 시뮬레이션할 수 없다는 것은 그들의 "인구 통계학적" 내부 표현이 피상적이거나, 훈련 데이터에 강력한 세그먼트 모델을 구축할 만큼 충분히 명시적이고 미묘한 인구 통계학적으로 연결된 선호도 데이터가 부족하다는 것을 시사한다.

이질성을 포착하지 못하는 것은 LLM이 방대한 집계된 훈련 데이터에서 파생된 "평균" 또는 "최빈값" 소비자 프로필에 내재적으로 편향되어 있음을 보여준다. LLM은 인구 통계학적 *레이블*을 모방할 수 있지만, 세그먼트별 선호도를 이끄는 문화적, 경제적, 심리적 요인의 복잡한 상호작용을 종합하는 데 어려움을 겪는다. 그들의 "인구 통계학적 이해"는 진정한 행동 모델링보다는 키워드 연관성에 가깝게 보인다. 따라서 세분화, 개인화된 마케팅 또는 틈새시장 이해에 중점을 둔 시장 조사자에게 LLM은 현재 불충분하다. 인간 중심의 연구 방법(설문조사, 민족지학, 포커스 그룹)은 이질적인 선호도를 밝히고 검증하는 데 여전히 필수적이다. LLM은 일반적인 시장 기준선을 제공할 수 있지만, 타겟 전략에 필요한 세분화된 통찰력은 제공하지 못한다.

이질성을 포착하지 못하는 것은 데이터 세분화 문제일 수 있다. LLM은 방대한 양의 텍스트를 가지고 있지만, 이 텍스트가 특정 선호도와 연결된 정확한 인구 통계학적 데이터로 일관되게 태그되어 있을 가능성은 낮다. 예를 들어, 리뷰는 제품을 언급할 수 있지만, 리뷰어의 소득, 성별, 그리고 이러한 요인이 특정 기능에 대한 WTP에 어떻게 구체적으로 영향을 미치는지 명시적으로 언급하는 경우는 드물다. 텍스트의 암시적인 인구 통계학적 신호는 LLM이 미묘한 세그먼트 차이를 안정적으로 학습하기에는 너무 약하거나 노이즈가 많을 수 있다. 이는 시장 조사에서 LLM의 근본적인 데이터 한계를 보여준다: LLM은 비정형 텍스트를 처리하는 데 탁월하지만, 원하는 결과가 훈련 코퍼스에 명시적이고 일관되게 존재하지 않는 정확하고 구조화된 인구 통계학적으로 연결된 선호도 데이터를 요구할 때는 어려움을 겪는다.

## **시장 조사자를 위한 실질적 시사점 및 권고**

### **LLM의 전략적 활용**

LLM은 시장 조사자에게 귀중한 도구로 활용될 수 있는 여러 전략적 영역을 제공한다.

* **초기 단계 기능 테스트 및 개선:** LLM은 광범위한 인간 설문조사에 투자하기 *전에* 새로운 기능 아이디어를 신속하게 테스트하고 개선하는 데 매우 유용하다.1 특히 유사한 맥락에서 이전 설문조사 데이터를 미세 조정에 활용할 수 있는 경우, 다양한 기능 구성에 대한 초기 WTP 추정치와 선호도 신호를 신속하게 제공할 수 있다.1 이는 제품 개발에서 더 빠른 반복을 가능하게 한다. 따라서 LLM을 제품 개발 수명 주기 내의 신속한 프로토타이핑 도구로 통합하여 기능 개념 및 잠재적 가격 전략을 신속하게 평가할 것을 권고한다.
* **인간 연구의 보완, 대체 아님:** LLM은 전통적인 인간 설문조사에 대한 비용 효율적인 보완책으로 작용할 수 있으며, 특히 인간 데이터가 부족하거나 수집 비용이 많이 드는 새로운 시장 환경에서 유용하다.1 LLM은 옵션을 좁히고, 개념의 우선순위를 정하며, 예비 시장 신호를 제공하여 인간 연구를 위한 자원 배분을 최적화하는 데 도움을 줄 수 있다.1 LLM을 초기 스크리닝, 광범위한 시장 통찰력 및 가설 생성에 사용하고, 후속 인간 연구는 검증, 심층 분석 및 미묘한 이해에 전념하는 하이브리드 연구 접근 방식을 채택하는 것이 바람직하다.
* **벤치마킹 및 "평균" 시장 통찰력:** LLM은 평균 인구 WTP 및 선택 점유율에 대한 빠른 벤치마크를 제공할 수 있다.1 연구는 LLM이 WTP와 같은 복잡한 관계보다 시장 점유율과 같은 더 간단한 관계를 더 쉽게 학습해야 하며, 미세 조정이 구매하지 않는 옵션에 대한 인간 선택 복제 능력과 선택 점유율의 유사성을 개선한다고 주장한다.1 이는 초기 시장 규모 측정이나 일반적인 소비자 정서 이해에 유용하다. LLM을 활용하여 시장의 전반적인 흐름을 파악하고, 제품 속성에 대한 일반적인 반응을 이해하며, 특정 인간 세그먼트와의 비교를 위한 기준선을 생성하는 것이 좋다.

### **통합을 위한 모범 사례 및 고려 사항**

* **미세 조정의 중요성:** 최적의 정확성과 인간 선호도와의 일치도를 위해, 특히 기존 제품 범주 내의 새로운 기능에 대해 LLM을 관련 과거 인간 설문조사 데이터로 미세 조정하는 것이 중요하다.1 이는 예측력을 크게 향상시킨다. 따라서 조직 내에서 과거 설문조사 데이터의 수집, 구조화 및 접근성을 우선시해야 한다. 이 데이터는 LLM을 효과적으로 활용하기 위한 전략적 자산이 된다.
* **신중한 프롬프트 엔지니어링:** LLM 응답의 품질과 특성은 프롬프트 문구에 크게 의존한다.1 연구자들은 원하는 유형의 응답을 유도하고, 내재된 편향을 최소화하며, 일관성을 보장하기 위해 프롬프트를 세심하게 설계해야 한다. 표준화된 프롬프트 템플릿을 개발하고 엄격하게 테스트해야 한다. LLM의 페르소나와 맥락을 안내하기 위해 다중 턴 대화 또는 역할극 프롬프트를 고려할 수 있다.
* **인간 데이터로 검증:** 특히 중요한 비즈니스 결정, 완전히 새로운 제품 범주를 탐색할 때, 또는 이질적인 선호도를 이해하는 것이 필수적일 때, LLM에서 파생된 중요한 통찰력은 항상 인간 연구로 검증해야 한다.1 LLM 통찰력이 인간 검증을 필요로 하는 시점에 대한 명확한 임계값과 기준을 설정해야 한다. 최종 전략적 결정에 LLM에만 의존하는 것을 피해야 한다.
* **LLM 한계 이해:** 새로운 제품 범주로의 확장 및 고객 이질성 정확한 포착에 대한 LLM의 현재 한계를 명확히 인지해야 한다.1 상당한 인간 감독, 전문가 해석 및 엄격한 검증 없이 이러한 특정 작업에 LLM을 사용해서는 안 된다. 시장 조사 워크플로우 내에서 LLM 유용성의 경계를 명확히 정의해야 한다. LLM이 명백히 저조한 성능을 보이는 영역에서는 과도한 의존을 피해야 한다.
* **LLM 진화 모니터링:** LLM 분야는 급속한 개발과 빈번한 새 모델 출시로 특징지어진다.1 연구자들은 효과적인 상태를 유지하기 위해 새로운 LLM 버전의 개선된 기능, 나타나는 한계 및 진화하는 모범 사례를 지속적으로 평가해야 한다. 최신 LLM 연구 및 모델 출시에 대한 정보를 계속 파악해야 한다. 조직과 관련된 특정 시장 조사 작업에 대한 새 모델의 성능을 평가하기 위해 내부 파일럿 연구를 수행해야 한다.

LLM의 강점(속도, 비용 효율성, 그럴듯한 WTP 생성 능력)과 약점(이질성, 새로운 범주)을 고려할 때, 그들의 가장 강력한 역할은 정교한 "가설 생성기"일 수 있다. 단순히 인간 연구를 대체하는 대신, LLM은 다양한 기능 조합에 대한 수많은 WTP 추정치를 신속하게 생성하고, 어떤 기능이 가장 유망하거나 문제가 되는지 식별할 수 있다. 그리고 이러한 *가설*은 인간 피험자로 엄격하게 테스트하고 검증될 수 있다. 이는 LLM의 역할을 "데이터 소스"에서 "연구 설계 가속기"로 전환시킨다. 시장 조사 워크플로우는 LLM을 초기 아이디어 구상 및 가설 수립 단계에서 사용하여 후속 인간 연구의 범위와 비용을 크게 줄임으로써 최적화될 수 있다. 이는 인간 연구자들이 광범위한 탐색에 시간을 덜 쓰고, 심층적인 검증과 미묘한 이해에 더 많은 시간을 할애할 수 있음을 의미한다.

내부 설문조사 데이터로 미세 조정이 성공했다는 점은 기업의 과거 시장 조사 데이터의 전략적 가치가 증가하고 있음을 보여준다. 종종 사일로화되거나 충분히 활용되지 않던 이 데이터는 LLM의 잠재력을 최대한 발휘하는 데 중요한 자산이 된다. 이는 일반적인 LLM을 특정 제품 라인 및 고객 기반에 대한 고도로 전문화된 "도메인 전문가"로 변모시킨다. 이는 경쟁 우위가 될 수 있다. 기업은 과거 시장 조사 데이터에 대한 강력한 데이터 거버넌스, 표준화 및 접근성에 투자해야 한다. 이 내부 데이터는 미세 조정에 사용될 때, 시장 정보에 LLM을 활용하는 데 있어 독점적인 우위를 제공하여, 일반 모델이 맞춤형 솔루션처럼 작동하도록 만들 수 있다.

## **한계점 및 향후 전망**

### **현재 한계점**

* **정적인 선호도:** LLM은 방대한 훈련 데이터에서 파생된 정적인 선호도를 드러낼 수 있으며, 이는 빠르게 진화하는 소비자 취향, 새로운 시장 트렌드 또는 역동적인 경쟁 환경을 완전히 포착하지 못할 수 있다.1 이러한 한계는 LLM 통찰력을 새로운 인간 연구로 보강함으로써 어느 정도 완화될 수 있다.
* **프롬프트 민감도:** LLM 응답의 품질, 미묘한 차이, 심지어 편향까지도 입력 프롬프트의 정확한 문구와 구조에 크게 의존한다.1 이는 신뢰할 수 있고 일관된 결과물을 보장하기 위해 세심한 프롬프트 엔지니어링과 광범위한 테스트를 필요로 한다.
* **빠른 개발 주기:** LLM 분야는 예외적으로 빠른 개발과 빈번한 새 모델 출시로 특징지어진다.1 이는 시장 조사자들이 효과적인 상태를 유지하기 위해 새로운 LLM 출시를 지속적으로 평가하고, 방법론을 조정하며, 모범 사례를 업데이트해야 함을 의미한다.
* **진정한 이해 부족:** 근본적으로 LLM은 진정한 인간의 이해, 직관, 감성 지능 또는 살아있는 경험을 가지고 있지 않다. 이러한 내재된 한계는 복잡한 감정적 동인, 매우 주관적인 선호도 또는 소비자 행동에 영향을 미치는 미묘한 사회문화적 뉘앙스를 포착하는 능력을 제한한다. 이러한 한계는 LLM이 패턴 인식 및 텍스트 생성을 위한 강력한 통계 도구이지, 의식 있는 존재가 아님을 강조한다. 그들의 결과물은 데이터의 패턴을 반영하는 것이지, 인간의 이해나 공감에서 비롯된 진정한 통찰력이 아니다.

LLM은 결과물(WTP 추정치)에 초점을 맞추지만, LLM이 특정 결과(예: MacBook에 대한 부정적인 WTP, "팬케이크 치약"에 대한 긍정적인 WTP)를 생성하는 *이유*를 깊이 파고들지는 않는다. 이러한 "블랙박스" 특성은 오류를 진단하거나 LLM의 근본적인 "논리"를 진정으로 이해하기 어렵게 만든다. LLM이 더 통합됨에 따라, 시장 조사에서 설명 가능한 AI(XAI)에 대한 수요가 증가할 것이며, 이는 연구자들이 모델을 더 효과적으로 신뢰하고 디버깅할 수 있도록 할 것이다. 따라서 시장 조사를 위한 LLM의 미래 연구 및 개발은 정확성뿐만 아니라 해석 가능성에도 초점을 맞춰야 한다. 시장 조사자들은 자신감을 구축하고 전략적 결정을 내리기 위해 LLM이 생성한 선호도의 *동인*을 이해하는 도구가 필요할 것이다.

### **향후 전망 및 낙관론**

연구 저자들은 LLM이 미래에 훨씬 더 정교해질 것이라는 낙관론을 표명하며, 이는 전통적인 인간 피험자 연구의 한계를 초월하는 시장 조사의 새로운 패러다임의 출현으로 이어질 수 있다고 본다.1 이는 더 미묘한 시뮬레이션, 복잡한 시나리오의 더 나은 처리, 그리고 다른 데이터 소스와의 개선된 통합을 포함할 수 있다.

이 연구의 예비적 성격은 상당한 미래 작업의 필요성을 강조한다. 여기에는 LLM 통합을 위한 더 강력한 모범 사례 식별, 현재의 한계(예: 이질성, 범주 간 확장)를 해결하기 위한 새로운 기술 개발, 그리고 시장 조사에서 LLM의 새롭고 혁신적인 응용 프로그램 탐색이 포함된다.

시장 조사의 미래는 점점 더 정교해지는 하이브리드 모델을 포함할 가능성이 높다. 이러한 모델에서 LLM은 대규모 데이터 합성, 초기 가설 생성 및 신속한 프로토타이핑을 효율적으로 처리하는 반면, 인간 연구자들은 질적 깊이, 전략적 해석, 복잡하거나 미묘한 통찰력 검증 및 윤리적 고려 사항 해결에 집중할 것이다.

LLM이 더욱 정교해지고 제품 개발 및 마케팅 전략에 잠재적으로 영향을 미치면서, 윤리적 고려 사항이 가장 중요해진다. LLM이 주로 "평균" 선호도를 반영하고 이질성에 어려움을 겪는다면, 기존 편향을 강화하거나 소수 소비자 그룹을 소홀히 할 위험이 있다. "정적인 선호도" 한계는 또한 LLM이 신선하고 다양한 데이터로 정기적으로 업데이트되지 않으면 구식 또는 해로운 고정관념을 영속화할 수 있다는 우려를 제기한다. LLM을 배포하는 시장 조사자들은 LLM 결과물의 잠재적 편향을 이해하고 완화할 윤리적 책임이 있다. 여기에는 훈련 데이터 다양성 보장, 다양한 인구 통계학적 그룹에 걸쳐 모델 성능 정기적으로 감사, 그리고 편향의 영속화 또는 특정 소비자 요구의 소홀을 방지하기 위한 인간 감독 유지가 포함된다. "새로운 패러다임"은 윤리적 AI의 토대 위에 구축되어야 한다.

#### 참고 자료

1. Using LLMs for Market Research - Harvard Business School, 7월 29, 2025에 액세스, <https://www.hbs.edu/ris/Publication%20Files/23-062_ed720ebc-ec4d-4bc3-a6ba-bad8cfbd9d51.pdf>