설문지를 받아든 AI '합성 페르소나',
소비자를 대신할 수 있을까?

글 | **이준원** 한국외대 언론학 박사 leejw34@gmail.com

지난 2월 스타트업 Simile가 1억 달러 규모의 시리즈 A 투자를 유치했다. 창업자는 가상 마을에 AI 에이전트를 풀어놓아 화제를 모았던 '생성 에이전트(Generative Agents)' 연구의 주역이다. 실제 사람의 인터뷰와 행동 데이터를 학습해 그 사람처럼 생각하고 답하는 '디지털 트윈(Digital Twin)'을 만들겠다는 이 회사에는 이미 CVS Health와 Gallup 등이 고객 및 파트너로 참여하며 가능성을 보여주고 있다.

그러나 곧이어 미국 매체 Axios가 AI 시뮬레이션으로 생성된 조사 결과를 실제 사람을 대상으로 한 여론조사처럼 인용했다가 뒤늦게 편집자 주를 추가하는 일이 발생했다. 뉴욕타임스에는 실리콘 샘플링이 여론조사를 망칠 수 있다는 학자들의 기고가 실렸고, 여론조사 업계에서도 우려의 목소리가 나왔다.

이처럼 같은 기술을 두고 실리콘밸리와 여론조사 업계는 정반대의 반응을 보이고 있다.
사람에게 직접 묻지 않고도 사람의 답을 얻겠다는 이 기술은
마케팅 리서치의 오랜 문법을 바꿀 혁신일까,
아니면 그럴듯한 착시에 불과할까.

응답자 없는 설문조사, 실리콘 샘플링의 등장

ChatGPT가 아직 출시를 앞둔 2022년 9월, Argyle과 동료들은 대규모 언어모델(LLM)에 실제 설문 응답자의 인구통계학적 배경을 1인칭 서사로 주입하면 해당 집단의 응답 분포를 상당히 정확하게 재현할 수 있다는 연구를 투고했다. 연구진은 이렇게 만들어진 가상의 응답자 집단을 '실리콘 샘플(Silicon Sample)'이라 불렀고, 이 방법이 작동하는 이유를 '알고리즘 충실도(Algorithmic Fidelity)'라는 개념으로 설명했다.

LLM은 방대한 텍스트를 학습하는 과정에서 인간 사회의 온갖 편향까지 함께 흡수하는데, 연구진은 LLM이 학습한 텍스트 안에 집단별 태도와 사회문화적 맥락의 차이가 일정 부분 반영되어 있으며, 적절한 조건을 부여하면 그러한 패턴을 불러낼 수 있다고 보았다. 제거해야 할 결함으로만 여겨지던 AI의 편향이, 조건만 정교하게 걸어주면 인간 집단의 다양성을 비추는 거울이 될 수 있다는 역발상이었다. 이후 경제학자 Horton이 LLM을 '호모 실리쿠스(Homo Silicus)'라 명명하며 고전 행동경제학 실험들을 재현하는 등 검증이 이어졌다.

이후 관련 기술은 목적과 재현 단위에 따라 여러 형태로 분화됐는데, 이는 하나의 진화 계보라기보다 서로 다른 조사 목적을 수행하는 도구들에 가깝다. 구체적으로는 실제 데이터를 원료로 부족한 표본을 늘리는 데이터 증강(부스팅), 프로필을 조건으로 걸어 집단의 정량 응답을 생성하는 실리콘 샘플링, 세그먼트를 인격화한 가상 인물과 대화하며 정성적으로 탐색하는 합성 페르소나, 그리고 특정 개인의 데이터를 학습해 그 사람의 후속 응답을 시뮬레이션하는 디지털 트윈으로 구분할 수 있으며, 각각의 간단한 특성은 표 1 과 같다.

시작되는 상품화

이러한 실험은 어느 새 리서치 업계의 신규상품이 되었다. Qualtrics는 지난해 합성 패널 서비스를 공개하며, 회사 측 추산으로 조사 비용을 최대 50% 줄이고 수주가 걸리던 결과 도출 시간을 수분까지 단축할 수 있다고 설명했다. 이 외에도 Ipsos가 스탠퍼드 대학과 손잡고 디지털 트윈 패널 구축에 나섰고, 앞서 언급한 Axios 사례의 주인공 Aaru의 기업가치가 10억 달러로 평가되는 등 업계의 본격적인 움직임이 감지되고 있다.

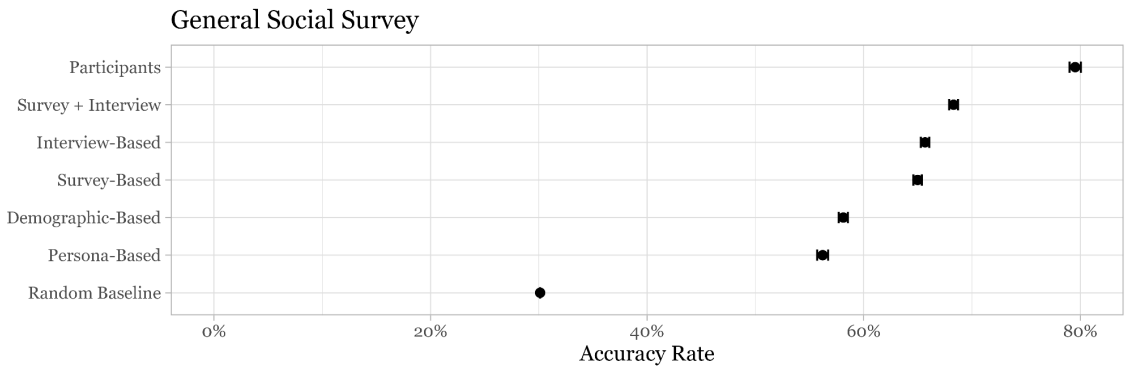
국내에서도 지난 4월 NVIDIA가 한국의 공식 통계를 기반으로 600만 건의 한국인 가상 페르소나 데이터셋을 상업적 이용까지 허용하는 조건으로 공개하며 기술적 진입 장벽을 낮췄고, 국내 리서치 기업들의 검증과 상용화 시도가 뒤따랐다. 오픈서베이는 합성패널의 검증 기준과 자체 실험 결과를 공개했고, 마크로밀 엠브레인은 6월 NVIDIA 데이터셋을 활용한 파일럿 결과를 발표하며 신제품 콘셉트나 네이밍 같은 자극물 평가에서는 합성 패널과 실제 조사의 차이가 상대적으로 작았지만, 브랜드 인지나 실제 소비 행동에서는 유의미한 차이가 확인되었다는 결론을 내놓았다.

장점과 한계

이러한 기술의 낙관을 뒷받침하는 근거는 분명히 존재한다. 한 연구는 미국인 1,052명을 각 2시간씩 심층 인터뷰한 뒤 개인별 생성 에이전트를 구축했다. 이 디지털 트윈들의 종합사회조사 문항 응답 일치율은 약 66%였다. 사람 역시 2주 뒤 자신의 답을 약 80%만 재현했다는 점을 감안하면, 인간의 자기 일관성을 기준으로 약 83% 수준의 성과였다. 사람의 기억조차 흔들리는 구간

구분	데이터 증강	실리콘 샘플링	합성 페르소나	디지털 트윈
핵심 목적	부족한 관측치 보완	집단의 정량 응답 생성	세그먼트의 동기와 언어 탐색	특정 개인의 응답 시뮬레이션
산출물	확장된 데이터셋	문항별 정량 응답	대화, 인터뷰, 서사	개인별 정량/정성 응답
적합 용도	희소 세그먼트 안정화	문항 파일럿, 초기 스크리닝	아이디에이션, 타깃 이해	재접촉, 반복 질의, 시나리오 탐색
위험 요소	원자료 편향 증폭	대표성 부족	고정관념 재생산	개인정보

표1 합성 소비자 접근법의 유형



까지 기계가 따라온 셈이다.

그러나 해당 결과는 2시간 분량의 실제 인터뷰로 에이전트를 실제 데이터에 뿌리내리게 하는, 이른바 '그라운딩(Grounding)'을 거쳤을 때의 이야기이며, 인구통계 프로필만 입력한 에이전트의 정확도는 눈에 띄게 낮았다. 또한, Bisbee와 동료들이 발표한 검증 연구에서 ChatGPT의 응답 평균은 실제 조사와 대체로 비슷했지만, 응답 분포는 인간 자료보다 지나치게 좁은 결과도 나타났다.

더 심각한 문제는 변수 간 관계였다. 합성 자료로 산출한 회귀계수의 절반 가까이가 실제 자료와 통계적으로 유의하게 달랐고, 그 가운데 약 32%는 효과의 방향까지 반대로 나타난 것이다. 평균의 유사성이 통계적 추론의 신뢰성을 보장하지 않는다는 뜻이다.

여기에 LLM의 심리적 프로필이 서구사회와 고학력 등 이른바 WEIRD 사회에 치우쳐 있으며 문화적 거리가 멀어질수록 인간 유사성이 급감한다는 점을 보였고, Gui와 Toubia는 가격만 달리한 가상의 실험에서 LLM이 경쟁 제품의 가격이나 소비자의 소득처럼 명시되지 않은 조건까지 함께 바꾸어 추론한다는 점을 보였다.

이상의 논의가 공통적으로 보여주는 것은 모델의 크기만으로는 부족하다는 점이다. 결국 성능을 좌우하는 것은 모델의 크기뿐 아니라, 어떤 실제 데이터에 얼마나 충실하게 뿌리내렸는가다.

90년 전 Literary Digest의 교훈

1922년 Lippmann은 그의 저서 '여론(Public Opinion)'에서 인간이 실재 그 자체가 아니라 머릿속에 구성된 세계, 곧 '의사환경(Pseudo-environment)'에 반응한다고 보았다. 이 관점에서 LLM이 생성하는 답은 여론 그 자체라기보다 인간이 텍스트로 남긴 여론의 흔적을 다시 조합한 결과다. 문제는 그것이 파생물이라는 사실이 아니라, 그 흔적에 누구의 목소리가 얼마나 담겼느냐에 있다.

이를 잘 보여준 것이 1936년 미국 대선이다. 잡지 Literary Digest는 전화번호부와 자동차 등록부 등에서 추린 1,000만 명에게 모의투표 용지를 발송했고, 237만 명의 응답을 모아 공화당 후보의 압승을 예측했다. 그러나 결과는 루스벨트의 역사적 압승이었다.

실패의 원인은 응답의 수가 아니라 조사에 포함되고 응답한 사람들의 구성이었다. 대공황기의 미국에서 전화와 자동차는 상대적으로 부유한 이들의 소유물이어서 명단 자체가 특정 계층으로 기울어 있었고, 24%에 그친 회신율이 낮은 비응답 편향까지 더해졌다. 반면 갤럽은 지역과 성별, 연령, 경제적 지위의 구성을 맞춘 할당표본으로 결과를 맞혔는데 그 규모는 고작 5만 명에 불과했다. 표본은 크기가 아니라 대표성이라는 교훈이 탄생한 순간이었다.

90년이 지난 지금 이 이야기가 다시 소환되어야 하는 이유가 있다. LLM의 학습 데이터는 전통적 의미의 표집틀은 아니지만, 어떤 사람들의 언어와 경험이 모델 안에 더 많이 남는지를 결정한다는 점에서 거대한 표집틀에 비유할 수 있다. 인터넷의 텍스트는 인류 전체가 아니라 온라인에서 활발히 글을 쓰는 특정 집단, 주로 서

구 영어 사용자들의 목소리를 과대 표집하며, 앞서 살펴본 WEIRD 편향 연구가 실증한 것이 바로 이 기울기다. 90년 전의 전화번호부가 부유층의 여론을 미국의 여론으로 착각하게 만들었듯, 아무런 보정 없이 모델에게 페르소나를 연기시키는 방식은 온라인 영어권의 그림을 인류의 그림으로 착각하게 만들 수 있다. 규모의 유혹에만 기댄다면, 합성 표본은 새로운 Literary Digest가 될 수밖에 없다.

대체가 아닌 보완, 그리고 응답의 자격

실리콘 샘플링과 디지털 트윈은 이미 상용화 단계에 들어섰지만, 아직 활용성이 충분히 검증된 것은 아니다. 지금까지의 연구를 종합하면 합성 응답자는 집단 수준의 경향을 빠르게 탐색하는 데에는 유용하지만, 실제 응답의 분산이나 변수 간 관계를 재현하고 통계적 추론과 예측에 활용하는 데에는 한계가 있다. 대표성은 응답의 자격을 얻기 위한 입장권일 뿐, 결과의 품질을 보증하는 졸업장이 아니다.

따라서 합성 응답자는 실제 조사를 대체하기보다 보완하는 도구로 활용해야 한다. 설문 문항의 사전 점검, 초기 콘셉트 스크리닝, 가설 탐색처럼 위험이 낮고 실제 자료와 비교하기 쉬운 영역에서 시작하고, 중요한 의사결정에는 인간 표본을 통한 별도의 검증을 거쳐야 한다. 합성 데이터의 사용 여부와 생성 방식, 인간 개입 범위, 검증 절차도 계약과 보고서에 명확히 밝혀야 한다. 아울러 과거 조사에 한 번 참여했다는 사실이 자신의 응답을 학습한 에이전트가 이후의 질문에 무한히 대신 답해도 된다는 뜻은 아니므로, 원자료의 수집 목적과 활용 범위, 학습 사용에 대한 동의와 삭제 요구 가능성까지 확인할 필요가 있다.

무엇보다 디지털 트윈의 시대에 가장 값진 자산은 역설적으로 기업이 직접 축적한 실제 고객의 목소리다. 실제 응답에 뿌리내리지 않은 합성 데이터는 원금 없이 불어나는 이자와 같으며, 좋은 고객 데이터를 축적한 기업일수록 더 신뢰할 수 있는 합성 소비자를 만들 수 있다. 결국 실리콘 소비자에게 필요한 것은 빠르게 답하는 능력이 아니라 누구를 반영하는지 설명하고, 실제 인간의 응답 분포와 관계를 재현하며, 그 과정이 투명하고 반복 가능하다는 것을 증명하는 일이다. 사람에게 직접 묻지 않는 조사가 늘어날수록 사람을 제대로 담아낸 데이터의 가치는 오히려 높아질 것이다. 실제 응답에 충분히 뿌리내리지 않은 합성 데이터는 신뢰할 만한 판단의 근거가 되기 어렵다.

필자 | 이준원
브랜드 커뮤니케이션을 전공하였으며 디지털 광고회사에서 근무한 후 현재 한국외대 미디어커뮤니케이션 연구소 및 미디어 외교센터 소속으로 디지털 미디어를 통한 커뮤니케이션 전략에 대한 연구를 진행하고 있다.