

한국형 합성 페르소나는 실제 응답을 재현하는가?: Nemotron-Personas-Korea 데이터셋을 활용한 실리콘 샘플링의 가능성과 한계

어광수¹⁾

요약

본 연구는 Nemotron-Personas-Korea를 활용하여 LLM 기반 실리콘 샘플링의 가능성과 한계를 검토하였다. 1,800개의 한국형 합성 페르소나에 Gemini 및 Claude 계열 4개 모델을 적용하고, 국가데이터처 사회조사(2025)의 기부 의향 및 유산기부 의향 응답과 비교하였다. 또한 270개 부표본을 활용하여 조건부 표현, 행동 구체성, 손실 프레임, 상징적 행위 등 문항 프레임링 효과를 분석하였다. 분석 결과, 모든 모델은 실제 조사보다 높은 기부 의향을 보였으며, 특히 유산기부 문항에서 과대추정이 크게 나타났다. 반면 문항 프레임링은 응답 분포에 강한 영향을 미쳤으며, 구체적 행동 의도를 묻는 문항은 가장 큰 응답 감소 효과를 보였다. 또한 동일한 페르소나에 대해서도 모델 간 응답 차이가 상당히 나타났다. 이러한 결과는 정교한 페르소나만으로는 합성-실제 간극이 완전히 해소되지 않으며, 문항 설계가 LLM 기반 사회 시뮬레이션 결과에 중요한 영향을 미칠 수 있음을 시사한다.

주요용어 : 실리콘 샘플링, 대규모 언어모델, 합성 페르소나, 프레임링 효과, 사회조사

1. 서론

행정 및 정책 연구에서 데이터 기반 분석에 대한 수요가 급증하는 가운데, 개인정보 보호법의 강화, 서버비 비용 증가, 응답률 하락 등으로 인해 전통적 서베이 방법론은 다양한 제약에 직면해 있다. 설문조사 한 건의 비용이 수백만 원에서 수천만 원에 달하고, 민감한 정책 문항일수록 응답 거부율도 높아지는 상황에서 대규모 언어모델(Large Language Model, 이하 LLM)을 활용하여 인간 집단의 응답을 시뮬레이션하려는 시도가 정치 여론 연구를 중심으로 확산되고 있다.

특히 Argyle et al.(2023)은 GPT-3에 인구사회학적 특성을 부여한 합성 응답자가 실제 미국 정치 여론조사 응답 분포를 근사적으로 재현할 수 있음을 보이며, 실리콘 샘플링(Silicon Sampling)이라는 새로운 연구 방법론을 제안하였다. 이러한 방법론을 활용한 연구는 경제학 실험, 사회학적 태도 측정, 가상 패널 구성 등으로 빠르게 확산되었다. 그러나 이와 동시에 합성 응답이 실제 응답 분포를 충실히 재현하지 못한다는 비판적 검증 연구들도 잇달아 발표되었다. 이들 연구는 LLM 응답의 문화적 편향, 낮은

1) 단독저자: 경기연구원 인구영향평가센터, 초빙연구위원, E-mail: egs@gri.re.kr

응답 다양성, 하위집단별 재현 격차 등을 보고하였다. 다만 이들 연구는 대체로 단편적 인구통계 정보만을 페르소나로 사용하였기 때문에, 합성 응답과 실제 응답 사이의 간극이 페르소나 묘사 부족의 문제인지, 아니면 LLM 자체의 구조적 편향에서 기인하는 것인지는 열린 질문으로 남아 있었다.

본 연구는 이 핵심 질문을 실증적으로 검증하고자 최근 공개된 Nemotron-Personas-Korea 데이터셋을 사용한다. Nemotron-Personas-Korea 데이터셋은 우리나라 인구 통계에 기반하여 직업, 가족, 문화적 배경이 유기적으로 결합된 다차원 한국형 합성 페르소나 데이터셋으로 이를 사용했음에도 합성과 실제의 간극이 좁혀지지 않는다면, 이는 페르소나 설계만으로 설명하기 어려운 모델 수준의 편향 가능성을 의심해 볼 수 있을 것이다. 나아가 본 연구는 합성과 실제의 간극을 줄이는 또 다른 변수로 문항의 프레이밍(Framing) 효과를 실증적으로 검증한다. 문항 설계 방법에 따라 합성과 실제의 간극이 어떻게 변화하는지를 분석하고자 한다.

본 연구의 학술적 기여는 세 가지로 요약할 수 있다. 첫째, 기존 연구들이 제기해 온 ‘합성과 실제의 간극이 단순한 페르소나 설계의 문제인가?’라는 질문을 한국형 고도화 페르소나를 활용하여 실증적으로 검증한다. 이를 통해 정교한 페르소나 데이터셋을 사용했을 때 합성 응답과 실제 응답 간 차이가 어느 정도 유지되는지를 확인함으로써, 페르소나 설계만으로 설명하기 어려운 모델 수준의 편향 가능성을 논의한다. 둘째, 프레이밍 문항을 적용한 실험을 통해 문항 설계가 합성-실제 간극에 미치는 영향을 분석하고, 가상 패널 기반 연구에서 활용할 수 있는 설문 설계에 대한 제언을 제공한다. 셋째, Nemotron-Personas-Korea 데이터를 사회과학 및 행정학 연구에 적용하여 대규모 합성 인구 데이터의 활용 가능성과 한계를 평가하고, 향후 행정 및 정책 연구에서 LLM 기반 가상 패널 활용 방안을 제언한다.

2. 이론적 배경 및 선행연구

2.1 실리콘 샘플링(silicon sampling)

LLM은 방대한 인간 언어 자료를 학습하는 과정에서 문장 생성 능력뿐 아니라 그 언어 자료에 내재된 사회적 규범, 가치관, 의사결정 양식까지 통계적 규칙성의 형태로 함께 학습한다. 이러한 특성에 주목하여 최근 연구들은 LLM을 개별 인간의 사고를 묘사하는 도구를 넘어 인간 집단의 행동 분포를 근사하는 계산적 모형으로 활용하는 방향을 모색하고 있다. 이 접근의 목적은 개별 응답자 한 명의 응답을 정확하게 예측하는 데 있지 않고, 충분한 수의 합성 응답을 생성하여 집단 수준의 응답 분포와 변수 간 관계 구조를 재현하는 데 있다. 이러한 문제의식에서 제안된 대표적인 방법론이 실리콘 샘플링이다.

Argyle et al.(2023)이 제안한 실리콘 샘플링은 LLM을 활용하여 인간 집단의 응답을 시뮬레이션하는 대표적인 방법론으로, 이후 다양한 가상 패널 연구의 출발점이 되었다. 이들은 ANES(American National Election Studies) 데이터를 활용한 분석에서 GPT-3

기반 합성 표본이 실제 응답과 높은 수준의 상관관계를 보이며 변수 간 복합적 관계 구조를 재현할 수 있음을 보여 실리콘 샘플링 연구의 발전에 기여하였다. 아울러 알고리즘이 실제 인간 집단의 응답 패턴을 얼마나 충실하게 재현하는지를 의미하는 알고리즘 충실도(algorithmic fidelity) 개념을 제안하였다. 나아가 실리콘 샘플링의 타당성을 평가하기 위해 사회과학적 튜링 테스트, 후방 연속성, 전방 연속성, 패턴 대응성 등 네 가지 평가 관점을 제시하였다. 아울러 실리콘 샘플링 실행 절차의 측면에서 세 단계로 정리될 수 있다. 첫째, 개별 응답자의 인구사회학적 특성이나 서술형 배경 정보를 프롬프트에 조건화(conditioning)하여 LLM에 가상의 응답자 역할, 즉 페르소나(persona)를 부여한다. 둘째, 실제 조사와 동일하거나 유사한 설문 문항을 제시하여 각 페르소나의 응답을 생성한다. 셋째, 생성된 응답을 집계하여 응답률·분포·집단 간 차이 등을 산출하고 이를 실제 조사 결과와 비교한다. 이때 비교의 단위는 개별 응답의 정확성이 아니라 집단 수준의 분포 재현성이며, 본 연구의 분석 설계 역시 이 절차를 따른다.

이러한 LLM 기반 시뮬레이션 연구는 크게 세 방향으로 나누어 볼 수 있다. 먼저 인간 행동 재현 가능성을 검증하는 연구로 Horton et al.(2026)은 LLM을 경제적 행위자(homo silicus)로 활용하여 고전 행동경제학 실험을 재현하였고, Park et al.(2024)은 인터뷰 기반 조건화를 통해 인간의 검사-재검사 신뢰도와 비교 가능한 수준의 예측 정확도를 검증하였다. 두 번째는 실리콘 샘플링 방법론 자체를 확장하는 연구로 Sun et al.(2024)은 집단 수준의 인구 분포만을 활용한 랜덤 실리콘 샘플링(Random Silicon Sampling)을 제안하였다. 세 번째는 위에서 정의한 페르소나 구축 기술 자체를 고도화하는 연구로, Ge et al.(2024)은 LLM 부트스트래핑 방식으로 10억 개의 합성 페르소나를 구축하는 PersonaHub를 제안하여 대규모 합성 페르소나 생성의 기술적 가능성을 보였다. 그러나 이러한 연구들은 대체로 페르소나의 정교화가 시뮬레이션 정확도를 향상시킬 것이라는 전제 위에서 수행되었으며, 정교한 다차원 페르소나를 사용하더라도 합성-실제 간극이 실제로 감소하는지에 대해서는 논의가 충분히 이루어지지 않았다.

2.2 페르소나 정교화 가설과 구조적 편향 가설

Argyle et al.(2023) 연구의 낙관적 결과 이후 다수의 비판적 검증 연구도 등장하였다. 먼저 Atari et al.(2023)은 다양한 국가의 심리측정 자료와 LLM의 응답을 비교한 결과, LLM의 응답이 WEIRD(Western, Educated, Industrialized, Rich, Democratic) 사회의 인간 응답과 가장 높은 유사성을 보이며 문화적 거리가 멀어질수록 예측 정확도가 크게 감소함을 확인하였다. 이는 LLM이 특정 문화권의 특성을 상대적으로 강하게 반영함을 시사한다. 응답 분포 자체의 편향도 보고되었다. Boelaert et al.(2025)은 LLM이 각 문항에서 강한 편향과 낮은 분산을 보이며 사회적 태도 공간의 제한된 일부 영역만을 반영하는 경향이 있음을 확인하였다. 또한 이러한 편향이 문항별로 상이한 방향으로 나타나는 패턴을 관찰하고 이를 '머신 편향(machine bias)'이라 명명하였다. 이들은 이러한 결과가 현재의 LLM이 의견 및 태도 연구에서 인간 응답자를 대체하는 데 한계가 있음을 시사한다고 주장하였다. Ma et al.(2025)은 GLES(German Longitudinal Election Study) 데이터를 분석하여 LLM이 특정 정치적 하위집단은 비교적 정확하게 재현

하지만 다른 집단은 그렇지 못함을 보여주었으며, 알고리즘 충실도가 정치적 하위집단에 따라 달라질 수 있음을 주장하였다. Bisbee et al.(2024)은 LLM 합성 응답의 다양성 부족 문제를 정량화하였다. 집단 평균은 실제 서베이와 유사하나 응답 분산이 현저히 낮아 소수 의견이 과소 대표되며, 회귀계수가 실제 데이터 추정치와 상당히 차이를 보였다. 그러나 이들 비판 연구들이 공통적으로 사용한 것은 성별, 연령, 교육 수준 등 단편적 인구통계학적 프롬프트였다. 이 점은 중요한 방법론적 반론의 여지를 남긴다. 만약 페르소나의 현실성이 부족한 것이 간극의 주요 원인이라면, 더 정교하고 다차원적인 페르소나를 사용할 경우 간극이 줄어들 것이라는 페르소나 정교화 가설(Persona Refinement Hypothesis)이 성립할 수 있다. 이 가설은 기존 비판 연구들이 직접 검증하지 않은 채 남겨둔 질문이었다.

페르소나 정교화 가설에 대한 대안적 설명으로, 합성-실제 간극이 페르소나 묘사 부족이 아니라 LLM 모델 자체의 구조적 특성에서 기인한다는 구조적 편향 가설이 있다. 이 관점의 핵심 메커니즘은 RLHF(Reinforcement Learning from Human Feedback, 이하 RLHF)이다. RLHF는 인간 평가자가 선호하는 응답을 생성하도록 LLM을 추가 훈련하는 과정으로, 이 과정에서 모델이 사회적으로 바람직하고 도덕적으로 긍정적인 방향으로 응답하는 편향이 내재화될 수 있다. Salecha et al.(2024)의 연구는 이러한 가설에 대한 중요한 실증적 근거를 제시하였다. 이들은 GPT-4, Claude 3, Llama 3, PaLM-2 등 주요 LLM이 Big Five 성격검사에서 사회적으로 바람직한 방향으로 응답하는 경향을 보이며, 그 편향의 크기가 인간 표준편차 기준 약 1.20σ (GPT-4)에 달함을 보고하였다. 또한 질문 순서 무작위화나 문항 페르프레이징 이후에도 이러한 패턴이 지속되었으며, 모델 간 편향 수준에는 차이가 있었으나 모든 모델에서 사회적 바람직성 편향이 관찰되었다. Xiao et al.(2024)은 RLHF가 다양한 인간 선호를 균형 있게 보존하기보다 일부 선호 분포가 과소대표될 수 있음을 분석하였다.

두 가설은 - ‘페르소나 정교화 가설’과 ‘구조적 편향 가설’ - 합성-실제 간극의 주요 원인을 서로 다르게 설명한다. 전자는 간극의 상당 부분이 페르소나 정보의 부족에서 비롯된다고 보며, 따라서 페르소나를 정교화할수록 예측 정확성이 향상될 것으로 예상한다. 반면 후자는 간극의 주요 원인이 모델 자체의 구조적 특성에 있다고 보며, 페르소나를 정교화하더라도 개선 효과가 제한적일 것으로 예측한다.

두 가설은 모두 기존 연구에서 부분적으로 지지를 받아왔으나, 정교한 다차원 페르소나를 활용하여 양자를 직접 비교한 연구는 드물다. 본 연구는 Nemotron-Personas-Korea 데이터셋을 활용하여 페르소나 정교화가 합성-실제 간극을 완화하는지 검증하고, 그 결과를 통해 두 설명의 상대적 타당성을 평가하고자 한다. 검증에는 국가데이터처 사회조사(2025)의 기부 의향 문항(Q1)과 유산 기부 의향 문항(Q2)을 사용한다. 각 문항의 전문과 선정 근거는 3.1절과 3.3절에 제시하였다.

2.3 문항 프레이밍 효과와 연구가설

합성-실제 간극이 발생하더라도, 문항의 프레이밍(Framing) 설계를 통해 그 크기를 조절할 수 있다는 관점이 주목받고 있다. 서베이 방법론의 고전적 연구(Schuman & Presser, 1996)에서 문항 표현 방식이 인간 응답에 영향을 미친다는 점은 오래전부터 알려져 있으나, LLM 가상 패널에서 이 효과가 얼마나 증폭되는지는 최근에야 본격적으로 연구되고 있다. Tjua et al.(2024)은 9개 LLM을 대상으로 조건부 표현이나 가설적 시나리오가 포함된 문항에서 LLM 응답이 인간보다 훨씬 더 긍정적으로 왜곡됨을 관찰하였으며, 이 효과가 RLHF를 거친 모델에서 더욱 두드러짐을 보였다.

그렇다면 인간 행태 이론에서 도출된 프레이밍 기법을 LLM에 적용하는 것이 타당한가? LLM은 인간이 생성한 대규모 텍스트 말뭉치(text corpus)로 훈련되며, RLHF를 통해 인간 평가자의 판단 기준을 학습한다. 이에 따라 일부 과업에서는 인간과 유사한 응답 패턴이 관찰되기도 한다(Tjua et al., 2024; Salecha et al., 2024; Horton et al., 2026). 따라서 인간 행태 이론에서 도출된 프레이밍 효과 역시 LLM 응답에서 일정 부분 관찰될 가능성을 가정해 볼 수 있다.

본 연구는 국가데이터처 사회조사(2025)의 유산 기부 의향 문항을 원문(Q2)으로 삼고, 이를 변형한 프레이밍 문항에 대한 응답 변화를 통해 서로 다른 이론에서 도출된 프레이밍 효과가 LLM에서도 나타나는지를 탐색적으로 검토하고자 한다. 각 문항의 전문은 3.3절 <표 3.1>에 제시하였으며, 이하에서는 네 가지 프레이밍 문항(Q2_A~Q2_D)에 대응하는 가설을 순서대로 제시한다.

첫째, Tjua et al.(2024)은 문항 표현 방식에 따라 LLM 응답이 체계적으로 변화할 수 있음을 보고하였다. 이는 설문 문항의 표현과 맥락이 응답에 영향을 미친다는 서베이 방법론의 고전적 논의와 맥을 같이하며(Schuman & Presser, 1996), 이러한 효과가 LLM 기반 응답에서도 나타날 가능성을 시사한다.

H1(조건부 표현 가설): 조건부 표현을 제거한 문항(Q2_A)은 조건부 표현을 포함한 원문(Q2)보다 낮은 동의율을 보일 것이다.

둘째, Ajzen(1991)의 계획행동이론(Theory of Planned Behavior)은 추상적 태도와 구체적 행동 의도를 구분하며, 지각된 행동 통제가 실제 행동 의도에 영향을 미침을 이론화하였다. LLM이 추상적 의향에 과도하게 동의하는 경향이 있다면, 구체적 행동 계획을 묻는 문항에서 현실적 제약이 더 강하게 반영될 수 있다.

H2(구체적 행동 의도 가설): 구체적 계획 문항(Q2_B)은 추상적 의향을 묻는 원문(Q2)보다 낮은 동의율을 보일 것이다.

셋째, Kahneman & Tversky(1979)의 전망이론(Prospect Theory)은 사람들이 객관적 결과가 아니라 준거점(reference point)을 기준으로 이익과 손실을 평가하며, 일반적으로 동일한 크기의 이익보다 손실에 더 민감하게 반응함을 설명한다. 이를 확장한 Tversky & Kahneman(1981)은 동일한 결과라도 이득으로 표현되느냐 손실로 표현되느냐에 따라 선택이 달라질 수 있음을 보여주었다. 이러한 프레이밍 효과는 인간의 의사결정 과정에서 널리 확인되어 왔으며, 인간이 생성한 대규모 텍스트를 학습한 LLM 역시 손실과 이득에 대한 상이한 언어적 패턴을 반영할 가능성이 있다. 따라서 손실을 강조하는

프레이밍은 LLM 응답에서 보다 신중한 판단을 유도하여 긍정 응답을 감소시킬 것으로 예상된다. 이로부터 다음의 가설을 설정한다.

H3(손실 프레임 가설): 손실 프레임을 적용한 문항(Q2_C)은 원문(Q2)보다 낮은 동의율을 보일 것이다.

넷째, 사회적 바람직성 편향(social desirability bias) 연구는 LLM이 질문 맥락에 포함된 규범적 신호에 영향을 받을 수 있음을 보여준다. Lee et al.(2024)은 GPT-4 기반 시뮬레이션에서 응답 전 응답 서약문(commitment statement)을 제시할 경우 사회적으로 바람직한 응답이 증가할 수 있음을 보고하였다. 이러한 결과는 LLM이 단순히 부여된 페르소나 특성만을 반영하는 것이 아니라 질문의 맥락과 규범적 기대에 민감하게 반응할 수 있음을 시사한다. 따라서 기부 의사와 관련된 문항에서도 질문이 내포하는 규범적 의미와 상징성이 응답에 영향을 미칠 가능성을 예상할 수 있다. 다만 Q2_D는 상징성을 강조하는 동시에 서약문 작성이라는 구체적인 행위를 묻는 형식을 취하므로, H2에서 논의한 행동 구체성 기제가 함께 작동하여 규범적 신호의 상방 효과를 상쇄하거나 역전시킬 가능성도 배제할 수 없다. 본 연구는 사회적 바람직성 이론의 예측에 따라 아래와 같이 가설을 설정하되, 검증 결과는 두 기제의 상대적 크기에 대한 탐색적 관찰로 해석한다.

H4(상징적 행위 가설): 상징성을 강조한 문항(Q2_D)은 원문(Q2)보다 높은 동의율을 보일 것이다.

이상의 네 가설은 각각 언어적 조건성(H1→Q2_A), 행동 구체성(H2→Q2_B), 인지적 준거점(H3→Q2_C), 상징적 행위(H4→Q2_D)라는 서로 다른 인지적 요소가 LLM 응답에 미치는 영향을 탐색하기 위해 설계되었다. 다만 각각의 프레이밍 문항이 하나의 인지적 요소만을 온전히 분리한다고 보기는 어려우므로, 본 연구는 각 가설을 인과적으로 검증하기보다 문항 표현의 변화가 응답 분포에 미치는 영향을 탐색적으로 관찰하는 데 목적을 둔다.

한편 프레이밍 실험을 Q2에만 적용한 것은 Q1이 이미 현재 시점의 직접적 기부 의향을 묻는 문항으로 구성되어 있어 조건부 표현 제거(H1)의 조작이 어렵고, H2~H4의 이론적 근거가 되는 가족 상속 및 유산기부 맥락 역시 Q2에만 존재하기 때문이다. 따라서 Q2는 동일한 기부 주제를 유지하면서도 표현 방식과 의미 부여에 따른 프레이밍 효과를 비교할 수 있는 적절한 실험 조건을 제공할 것으로 예상된다.

2.4 연구문제 및 분석틀

지금까지의 논의를 종합하면, LLM 기반 실리콘 샘플링의 정확성을 둘러싼 설명은 크게 두 가지로 구분된다. 하나는 페르소나 정보의 부족이 합성-실제 간극의 주요 원인이라는 페르소나 정교화 가설이며, 다른 하나는 RLHF 등 모델 수준의 구조적 편향이 주요 원인이라는 구조적 편향 가설이다. 앞의 논의를 바탕으로 본 연구는 Nemotron-Personas-Korea의 다차원 한국형 페르소나 데이터셋을 활용하여 다음 두 연구문제를 검증한다.

RQ1. 다차원 한국형 페르소나를 활용할 경우 합성-실제 간극은 감소하는가?

RQ2. 프레이밍 문항은 합성-실제 간극을 완화하거나 강화하는가?

RQ1은 합성-실제 간극의 원인이 페르소나 정보의 부족에 있는지, 혹은 모델 수준의 구조적 편향에 있는지를 검토하기 위한 것이다. RQ2는 문항 프레이밍이 LLM 응답에 미치는 영향을 분석함으로써, 프레이밍 설계가 합성-실제 간극을 완화하거나 강화하는지를 탐색적으로 평가하기 위한 것이다. 이를 위해 앞서 논의한 조건부 표현 제거, 행동 구체화, 손실 프레이밍, 상징적 행위에 관한 네 가지 가설을 설정하였다.

3. 연구 방법

3.1 분석 데이터

본 연구는 앞서 논의한 연구문제를 검증하기 위하여 두 가지 데이터를 사용하였다. LLM 기반 합성 응답과 실제 조사 응답을 비교하기 위하여 합성 응답 생성에는 Nemotron-Personas-Korea 데이터셋을 사용하였으며, 실제 응답과의 비교를 위한 데이터로는 국가데이터처 사회조사(2025) 마이크로데이터를 활용하였다.

먼저 Nemotron-Personas-Korea 데이터셋은 NVIDIA가 2026년 4월 공개한 대규모 한국형 합성 페르소나 데이터셋이다. 이 데이터셋은 한국어 기반 AI와 에이전트 시스템이 국내 인구구조와 문화적 맥락을 재현하는 것을 목표로 구축되었다. 이를 위해 국가데이터처(KOSIS), 대법원, 국민건강보험공단 등 공공 통계자료를 활용하여 실제 한국 사회의 인구학적, 지리적 특성 분포를 재현하였으며, 성별, 연령, 교육수준, 직업, 혼인 상태, 거주지역 등의 속성과 함께 생활양식, 가치관, 관심사 등을 포함한 서술형 페르소나를 제공한다. 따라서 기존 실리콘 샘플링 연구에서 주로 사용된 단순 인구통계 프로파일보다 훨씬 풍부한 개인 맥락 정보를 제공한다는 점에서 페르소나 정교화 가설을 검증하기에 적합한 자료라 할 수 있다.

합성 응답과 비교할 실제 데이터는 국가데이터처 사회조사(2025) 마이크로데이터를 활용하였다. 사회조사는 전국 가구와 가구원을 대상으로 복지, 노동, 사회참여, 문화 및 삶의 질에 관한 인식과 행태를 조사하는 국가승인통계로 우리나라 국민의 사회적 태도와 가치관을 대표성 있게 파악하기 위해 매년 실시되고 있다. 본 연구에서는 만 20세 이상 성인 응답자 31,818명을 분석 대상으로 사용하였으며, 분석 시 가구원 가중치를 적용하였다.

분석 문항으로는 사회조사에 포함된 기부 관련 문항 2개(기부 의향, 유산 기부 의향)를 활용하였다. 두 문항은 현재의 기부 의향과 미래의 유산 기부 의향이라는 서로 다른 유형의 친사회적 행동을 측정한다는 점에서, LLM 기반 실리콘 샘플링의 재현성과 프레이밍 효과를 검증하기에 적합한 문항으로 판단하였다.

3.2 분석 모델 선정

본 연구는 Google과 Anthropic이 제공하는 4개의 LLM을 분석 모델로 선정하였다. 고성능 모델로는 Gemini 2.5 Pro와 Claude Sonnet 4.5를 경량 모델로는 Gemini 2.5 Flash와 Claude Haiku 4.5를 활용하였다. 이는 운영사(Google, Anthropic)와 성능 티어(고성능, 경량)의 2×2 교차 비교를 위한 설계이며, 합성-실제 간극이 특정 기업의 모델에 국한된 현상인지, 또는 서로 다른 계열의 LLM에서 공통적으로 나타나는 현상인지를 비교할 수 있도록 설계하였다.

아울러 모델 선정의 근거는 세 가지이다. 첫째, 최근 LLM 연구는 특정 모델의 성능 자체보다 모델 간 일관되게 나타나는 응답 패턴에 주목하고 있다. Salecha et al.(2024)은 GPT-4, Claude 3, Llama 3, PaLM-2 등 상이한 계열의 모델에서 사회적 바람직성 편향(Social Desirability Bias)이 반복적으로 관찰됨을 보고하였다. 따라서 단일 모델만을 활용할 경우 특정 모델의 특성이 일반적 현상으로 과대 해석될 위험이 있으며, 서로 다른 제조사의 모델을 비교하는 것이 필요하다.

둘째, Google과 Anthropic은 현재 가장 널리 활용되는 범용 LLM 모델로, 학계와 산업계에서 모두 높은 활용도를 보이고 있다. 두 기업은 모델 구조와 정렬(alignment) 전략에서 차이를 보이고 있어, 합성-실제 간극이 특정 정렬 방식의 결과인지 확인하는 데 적절한 비교 집단을 제공할 것으로 판단하였다.

셋째, 고성능 모델과 경량 모델의 비교는 모델 규모와 추론 능력이 합성 응답의 정확도에 미치는 영향을 탐색할 수 있는 기회를 제공한다. 특히 Gemini 2.5 Pro와 Gemini 2.5 Flash는 API 기본 설정에서 내부 추론(thinking) 과정을 활성화한 상태로 응답을 생성하는 반면, Claude Sonnet 4.5와 Claude Haiku 4.5는 기본 설정에서 별도의 추론 과정을 명시적으로 수행하지 않는다. 따라서 본 연구는 내부 추론이 활성화된 Gemini 계열과 그렇지 않은 Claude 계열의 비교를 통해, 추론 연산이 실제 응답 분포의 재현성 향상이나 RLHF 기반 정렬 편향 완화에 기여하는지를 탐색적으로 검토할 수 있다. 다만 본 설계에서는 내부 추론의 활성화 여부와 개발사가 서로 일치하므로 두 요인의 효과를 분리하기는 어려우며, 이에 관련 해석은 탐색적 관찰로 제한한다.

아울러 ChatGPT 계열 모델은 연구 설계 당시 API 비용과 대규모 반복 실험의 현실적 제약으로 인해 분석 모델로 선정하지 못했다. 다만 본 연구는 서로 다른 제조사와 성능 티어를 포함한 비교 설계를 통해 특정 모델이 아닌 LLM 전반의 응답 패턴을 검토하는 데 초점을 두었다.

3.3 검증 문항 및 프레이밍 문항

앞서 언급한 바와 같이 검증 문항으로는 국가데이터처 사회조사(2025)에 포함된 기부 관련 문항 2개를 선정하였다. Q1은 향후 기부 의향을 측정하는 문항으로, “귀하는 향후 1년 이내에 현금 또는 물품을 기부하실 의사가 있으십니까?”이며, 실제 조사에서의 ‘예’라고 응답한 비율은 40.1%였다. Q2는 유산 기부 의향을 측정하는 문항으로, “귀하는 향후 여건이 허락된다면 유산의 일부를 사회에 기부하실 의사가 있으십니까?”이며 실제

조사에서의 ‘예’라고 응답한 비율은 21.7%였다.

이 두 문항의 선정은 본 연구의 방법론적 설계와 밀접하게 관련된다. 두 문항 모두 사회적 기여와 나눔이라는 규범적 가치를 다룬다는 점에서 유사하지만, 실제 행동에 수반되는 비용과 제약 수준은 상당히 다르다. Q1은 비교적 적은 비용으로 단기간 내 실천 가능한 기부 의향을 묻는 반면, Q2는 사후 재산 처분과 가족 상속 문제를 포함하는 고비용 의사결정을 전제로 한다. 따라서 두 문항을 비교함으로써 사회적 바람직성에 대한 일반적 반응과 실제 행동 비용이 높은 상황에서의 응답 패턴을 구분하여 검토할 수 있다. 또한 Q2는 ‘여건이 허락된다면’이라는 조건부 표현을 포함하고 있어, 프레이밍 효과를 검증하기 위한 기준 문항으로 활용하기에 적합할 것으로 판단하였다.

프레이밍 효과를 검증하기 위하여 본 연구는 유산기부 의향 문항(Q2)을 기준으로 네 가지 변형 문항을 추가 구성하였다. 각 문항은 이론적 배경에서 제시한 조건부 표현 제거(H1), 구체적 행동 의도(H2), 손실 프레임(H3), 상징적 행위(H4)를 반영하도록 설계하였다. 원문 문항은 “귀하는 향후 여건이 허락된다면 유산의 일부를 사회에 기부하실 의사가 있으십니까?”이며, 이를 기준으로 조건부 표현 제거(Q2_A), 행동 계획 구체화(Q2_B), 손실 프레임 적용(Q2_C), 상징적 행위(Q2_D) 문항을 구성하였다.

프레이밍 실험을 Q2에 한정된 것은 연구가설과 문항 특성 간의 적합성을 고려한 것이다. Q2는 원문 자체에 조건부 표현을 포함하고 있으며 유산 상속이라는 미래 의사결정 맥락을 전제로 하고 있어, 조건부 표현(H1), 행동 구체화(H2), 손실 프레임(H3), 상징적 행위(H4)와 관련된 프레이밍 조작을 일관되게 적용할 수 있다. 반면 Q1은 일반적인 기부 의향을 묻는 문항으로 이러한 맥락적 요소를 포함하지 않으므로, 동일한 프레이밍을 적용할 경우 문항의 개념적 의미가 달라질 수 있다. 따라서 본 연구에서는 Q2를 프레이밍 실험의 대상으로 선정하고, Q1은 원문 비교를 위한 기준 문항으로 활용하였다.

이를 통해 본 연구는 동일한 정책 주제와 응답 맥락을 유지한 상태에서 프레이밍 방식만을 변화시켜, 문항 설계가 LLM 응답과 합성-실제 간극에 미치는 영향을 비교하고 분석한다.

<표 3.1> 프레이밍 문항

버전	문항	이론적 근거
원문	향후 여건이 허락된다면 유산의 일부를 사회에 기부하실 의사가 있으십니까?	사회조사 원문 기준, 조건부 표현 포함
Q2_A	실제로 유산의 일부를 사회에 기부하실 의사가 있으십니까?	조건부 표현 제거, 간극의 원인이 조건부 표현인지 검증
Q2_B	현재 유산의 일부를 사회에 기부할 구체적인 계획이 있으십니까?	추상적 의향→구체적 행동 의도로 전환
Q2_C	귀하 사후 가족이 아닌 사회에 재산 일부가 돌아가는 것에 동의하십니까?	이익 프레임→손실 프레임 전환
Q2_D	사망 후 재산 일부를 사회에 환원하는 유언장을 작성할 의향이 있으십니까?	유산기부의 상징성을 강조한 프레이밍

3.4 검증 설계 및 분석 방법

본 연구의 검증과 분석은 세 부분으로 구성된다.

첫째, 분포 적합성 검증에서는 Nemotron-Personas-Korea 데이터셋의 인구통계 변수 분포가 국가데이터처 사회조사(2025)의 응답자 분포와 얼마나 유사하게 구성되어 있는지를 분석하였다. 분포 간 차이는 전체 변동 거리(Total Variation Distance, TVD)를 활용하여 평가하였으며, TVD는 두 확률분포 간 차이를 0과 1 사이의 값으로 나타내며 0에 가까울수록 분포 간 유사성이 높은 것으로 해석된다.

둘째, 합성-실제 응답 분포 비교에서는 성별(2)×연령대(3)×지역(3)의 18개 집단 조합으로부터 각 100개씩 층화 무작위 추출한 총 1,800개의 페르소나에 대해 사회조사 문항 원문(Q1, Q2)을 4개의 언어모델에 적용하였다. 응답 생성에는 각 모델의 API 기본 설정을 사용하였으며(temperature 1.0, top-p 기본값), 페르소나와 문항 조합별로 각 1회씩 응답을 생성하였다. 이는 개별 페르소나의 결정론적 최빈 응답이 아니라 페르소나 집단의 응답 분포를 재현하는 데 목적을 둔 설계에 따른 것으로, temperature 1.0에서 각 응답은 모델 응답 분포로부터의 무작위 추출로 간주할 수 있다고 판단하였으며, 집계 수준의 응답률은 해당 분포의 불편추정치가 된다. 모델 스냅샷, 프롬프트 템플릿, 응답 코딩 규칙 등 세부 정보는 부록에 제시하였다.

다만 분석 표본은 집단 간 응답 패턴 비교를 위해 18개 집단에 균등 배분(집단별 100명)하여 구성하였으므로, 합성 표본의 단순 평균은 실제 인구 구성과 다른 가상의 모집단을 대변한다. 따라서 전체 응답률을 실제 조사와 비교할 때에는 2025년 사회조사 가구원가중값으로부터 산출한 집단별 인구 비중을 적용한 사후층화 가중 평균을 함께 제시한다. 또한 실제 사회조사는 복합표본설계에 기반하나 공개 마이크로데이터의 제약상 이를 온전히 반영한 분산추정에는 한계가 있다. 이에 본 연구의 합성-실제 비교는 엄밀한 통계적 검정이 아니라 실제 조사값을 기준으로 한 기술적 비교(descriptive comparison)로 제한하며, 제시하는 신뢰구간은 합성 표본의 이항 변동만을 반영한다.

이러한 제약을 고려하여, 각 모델의 전체 응답률과 실제 사회조사 응답률의 차이는 유의성 검정보다 간극의 크기와 효과크기(Cohen's d)를 중심으로 제시하였다. 또한 집단 수준의 분포 재현성을 평가하기 위해 성별·연령·지역 하위 집단별 절대오차(MAE, RMSE)와 집단 간 응답률 상관(Pearson r , Spearman ρ)을 함께 산출하였다.

셋째, 프레이밍 효과 분석에서는 합성-실제 응답 비교 분석에 사용된 표본 중 집단별로 15개씩 층화 무작위 추출한 270개의 부표본을 활용하였다. 해당 표본에 대해 원문을 포함한 5가지 문항 프레이밍 버전(Q2, Q2_A~Q2_D)을 4개의 언어모델에 적용하여, 문항 표현 변화가 응답 분포에 미치는 영향을 분석하였다. 프레이밍 실험은 동일한 페르소나에 대해 문항 표현만을 변경한 반복 측정 구조를 가지므로, 각 프레이밍 조건과 원문 조건 간 응답률 차이는 대응 표본 구조를 반영하는 정확 McNemar 검정으로 비교하였다.

한편 temperature 1.0으로 생성된 응답은 실행마다 달라질 수 있는 확률적 산출물이므로, 프레이밍 반복률(43절)과 개별 응답 쌍을 비교하는 모델 간 응답 일치율(4.4절)의 해석에는 생성 자체의 확률성(within-model variance)이 개입할 수 있다. 이를 통제하기 위해 프레이밍 표본 270명 전원을 대상으로 Q1을 포함한 6개 문항(Q1, Q2 원문, Q2_

A~Q2_D)을 모델별로 3회 반복 생성하는 검증 실험을 별도로 수행하였다(총 19,440회). 이로부터 산출한 모델별 자기 일치율을 모델 간 일치율 해석의 기저선으로 사용하였으며, 프레이밍 번복률에 대해서는 관찰 번복률에서 각 모델의 자기 불일치율을 차감한 초과 번복률을 함께 제시하여 생성 확률성의 영향을 분리하고자 하였다. 아울러 이 반복 검증이 본실험으로부터 일정 기간 경과 후 수행된 점을 이용하여, 원실험 응답과 반복 생성 응답 간의 시점 간 안정성도 함께 확인하였다. 반복 검증의 생성 설정과 프롬프트, 응답 코딩 규칙은 본실험과 동일하며, 세부 사항과 호출 시점은 부록에 제시하였다.

본 연구의 분석은 응답 간 차이의 통계적 유의성을 입증하는 데 목적이 있는 것이 아니라, 합성 응답이 실제 응답의 분포 및 집단 간 관계 구조와 얼마나 유사하게 정렬되는지를 평가하고, 문항 프레이밍 변화에 따른 응답 패턴이 모델 및 모델 계열 간에 얼마나 일관되게 나타나는지를 탐색적으로 확인하는 데 목적이 있다. 따라서 결과 해석은 유의확률(p-value)이 아니라 효과크기와 분포적 차이를 중심으로 이루어졌다.

4. 분석 결과

4.1 Nemotron-Personas-Korea 데이터셋 전체 분포 검증

본 연구에서는 먼저 Nemotron-Personas-Korea의 인구통계 변수 분포가 국가데이터처 사회조사(2025)의 응답자 분포와 얼마나 유사한지를 검증하였다. 사회조사는 가중치 적용 시 우리나라 인구를 대표하도록 설계된 국가승인통계이므로, 본 연구는 가구원 가중값을 적용한 사회조사 분포를 비교 기준으로 사용하였다. 검증은 두 수준에서 이루어졌다. 하나는 페르소나 데이터셋 전체(100만 개)의 분포이고, 다른 하나는 이후 응답 비교에 사용되는 1,800명 분석 표본의 분포이다. 분석 표본은 집단 간 비교 가능성과 통계적 균형을 확보하기 위해 성별(2)×연령대(3)×지역권역(3)의 18개 집단에서 각 100명씩 층화 균등추출한 것으로, 모집단의 실제 분포를 그대로 재현하기보다 집단 간 균형을 우선한 설계이다. 분포 간 차이는 전체 변동 거리(이하 TVD)로 측정하였는데, TVD는 두 분포가 각 범주에 부여하는 비율 차이의 절댓값을 모두 더한 뒤 절반을 취한 값으로, 0에 가까울수록 두 분포가 유사하고 1에 가까울수록 다름을 뜻한다. 본 연구는 이러한 TVD 값이 0에 가까운 정도를 기준으로 분포의 재현 수준을 판단하되, 변수 간 상대적 비교를 함께 고려하여 어떤 변수에서 편차가 상대적으로 크게 나타나는지를 해석하였다. 두 수준의 TVD를 함께 제시한 결과는 <표 4.1>과 같다.

성별, 지역, 연령은 전체 데이터셋에서 사회조사 분포를 매우 높은 수준으로 재현하였다. 성별과 지역의 TVD는 각각 0.001과 0.005 수준으로 사실상 사회조사 분포와 거의 동일하였으며, 연령 또한 TVD 0.003으로 모든 연령집단에서 차이가 0.3%p 이내 머물렀다. 분석 표본에서도 성별(TVD 0.005)과 지역(TVD 0.007)은 동일하게 높은 적합성을 유지하였다. 다만 연령대는 분석 표본에서 TVD 0.043으로 다소 커졌는데, 이는 각 연령집단을 동일 규모로 추출한 균등추출 설계의 결과이며 데이터셋 자체의 편차가 아니다.

반면 학력과 혼인상태에서는 제한적인 수준의 편차가 확인되었다. 전체 데이터셋에서 학력은 TVD 0.072로 중졸 이하와 고졸 비중이 사회조사보다 높고 전문대졸 및 4년제 대졸 비중은 낮게 나타나, 고학력 집단이 상대적으로 과소 재현되는 경향을 보였다. 혼인상태는 TVD 0.049로, 미혼·사별·이혼이 다소 높고 배우자있음이 낮게 나타났다. 또한 사회조사에 포함된 별거 범주가 데이터셋에는 존재하지 않아 범주 구성상의 차이도 확인되었다. 이러한 경향은 분석 표본에서도 유사하게 유지되었다(학력 TVD 0.049, 혼인상태 TVD 0.047). 한편 본 절의 TVD는 범주 구성의 차이 자체를 확인하기 위해 별거를 독립 범주로 유지한 상태에서 산출하였다. 반면 4.2절의 반복비례가중은 두 자료의 범주가 일치해야 하므로, 법적 혼인이 유지되는 상태임을 고려하여 별거를 배우자있음에 포함하였다(<표 4.4> 주2).

주거형태는 분석 표본에서 편차가 가장 크게 나타났다(전체 TVD 0.081, 표본 TVD 0.113). 전체 데이터셋을 기준으로 단독주택 거주 비율이 사회조사보다 약 8%p 낮게 나타난 반면, 아파트 비율은 사회조사와 유사한 수준이었다. 연립·다세대와 기타 거처는 사회조사보다 다소 높게 재현되었다. 이러한 경향은 전체 데이터셋과 분석 표본에서 공통적으로 관찰되었다. 다만 전체 데이터셋 수준에서는 주거형태(0.081)와 학력(0.072)의 TVD가 유사하여, 두 변수의 편차 크기를 구분하여 해석하기는 어렵다.

Nemotron-Personas-Korea는 성별·연령·지역과 같은 기본 인구통계 특성에 대해서는 전체 데이터셋과 분석 표본 모두에서 사회조사 분포를 상당 부분 재현하고 있는 것으로 판단된다. 반면 학력·혼인상태·주거형태는 성별·연령·지역에 비해 상대적으로 큰 편차가 확인되었다. 이 중 주거형태는 분석 표본에서 편차가 가장 컸으며, 단독주택이 과소 재현되었다. 이러한 편차는 이후 응답 결과를 해석할 때 고려해야 할 요인이므로, 학력·주거형태·혼인상태의 분포 차이가 관찰된 간극에 미치는 영향은 4.2절의 민감도 분석에서 별도로 검토한다.

<표 4.1> Nemotron-Personas-Korea 분포 검증: 전체 데이터셋과 분석 표본

변수	전체 TVD	표본 TVD	사회조사 대비 차이
성별	0.001	0.005	-
지역	0.005	0.007	-
연령대	0.003	0.043	전체는 모든 집단 0.3%p 이내, 표본은 균등추출 설계
학력	0.072	0.049	중졸 이하·고졸 과다, 전문대졸·4년제 대졸 과소
혼인상태	0.049	0.047	미혼·사별·이혼 과다, 배우자있음 과소, 별거 범주 없음
주거형태	0.081	0.113	단독주택 과소, 연립·다세대·기타 과다

4.2 합성-실제 응답 간극 분석

먼저 각 모델의 전체 응답률을 실제 사회조사 응답률과 비교하였다. 분석 표본은 18개 집단에 균등 배분되어 있어 단순 평균이 실제 인구 구성과 다른 가상 모집단을

대변하므로, 실제 응답률과 비교할 때에는 사회조사(2025) 가구원가중값 기반 성별×연령×지역 18개 집단의 인구 비중을 적용한 사후층화 가중 평균을 함께 산출하였다. <표 4.2>에는 단순 평균과 가중 평균을 병기하였으며, 실제 응답률과의 간극 및 효과크기(Cohen's h)는 가중 평균을 기준으로 산출하였다. 신뢰구간은 성별×연령×지역 18개 집단의 사후층화 가중 평균에 대응하는 층화 95% 구간으로, 합성 표본의 이항 변동만을 반영하며 사회조사의 복합표본설계에 따른 분산은 반영하지 않았다. 효과크기 Cohen's h 는 0.2, 0.5, 0.8을 각각 작은, 중간, 큰 효과의 관례적 기준으로 해석하였다(Cohen, 1988). 분석 결과 가중 적용에 따른 응답률 변화는 모든 모델과 문항에서 최대 1.6%p 이내로 단순 평균과 실질적 차이가 거의 없었다. 이는 관찰된 간극이 균등추출 설계에서 비롯되었을 가능성이 낮음을 시사한다. 다만 가중 적용의 방향은 모델에 따라 달랐다. Claude Sonnet 4.5는 가중 후 응답률이 소폭 낮아진 반면 나머지 모델은 높아졌는데, 이는 집단별 응답 패턴이 모델마다 다르기 때문으로 보인다.

간극의 크기는 문항에 따라 달랐다. Q1(기부 의향)에서는 Gemini 2.5 Pro가 +3.8%p($h=0.08$)로 실제 응답에 가장 근접하였으나, 나머지 세 모델은 실제보다 +18%p 이상 높은 동의율을 보였고 Claude Haiku 4.5는 +51.5%p($h=1.18$)에 달하였다. Q2(유산 기부 의향)에서는 모든 모델에서 간극이 확대되어, 실제 대비 최소 +23.9%p에서 최대 +65.9%p 더 높은 동의율이 나타났다. 특히 Gemini 2.5 Pro는 Q1에서 작은 효과에도 못 미치던 간극($h=0.08$)이 Q2에서는 중간 효과를 상회하는 수준($h=0.61$)으로 커져 간극이 확대되었다. Q1과 Q2 모두에서 실제 사회조사 응답의 동의율보다 합성 응답의 동의율이 높게 나타났다. 실제 사회조사에서 기부 의향(40.1%)이 유산 기부 의향(21.7%)보다 동의율이 18.4%p 높았던 반면, 합성 응답에서는 이러한 문항 간 격차가 모델별로 축소되거나(Claude 계열, 4.0~12.5%p) 역전되는(Gemini 계열, Q2가 Q1보다 최대 6.5%p 높음) 양상을 보여, 실제 응답자가 두 문항을 구분하는 정도를 합성 응답이 재현하지 못하였다.

<표 4.2> 전체 응답률: 단순·가중 평균과 효과크기

모델	문항	단순 평균	가중 평균 [95% CI]	실제 조사	간극	Cohen's h
Gemini 2.5 Pro	Q1	43.2%	43.9% [41.5, 46.5]	40.1%	+3.8%p	0.08
	Q2	50.1%	50.4% [47.9, 52.9]	21.7%	+28.7%p	0.61
Gemini 2.5 Flash	Q1	56.7%	58.3% [55.8, 60.7]	40.1%	+18.2%p	0.36
	Q2	58.3%	59.2% [56.7, 61.7]	21.7%	+37.5%p	0.79
Claude Sonnet 4.5	Q1	58.9%	58.1% [55.5, 60.6]	40.1%	+18.0%p	0.36
	Q2	46.6%	45.6% [43.2, 48.1]	21.7%	+23.9%p	0.51
Claude Haiku 4.5	Q1	91.2%	91.6% [90.1, 92.8]	40.1%	+51.5%p	1.18
	Q2	87.5%	87.6% [85.8, 89.2]	21.7%	+65.9%p	1.45

주: $n=1,800$. 간극과 Cohen's h 는 가중평균 기준

다음으로 개별 응답률 비교에서 나아가 인구집단 간 응답 구조가 얼마나 재현되는지를 평가하였다. 성별×연령×지역 18개 집단별로 합성 응답률과 실제 응답률을 대응시켜 집단 수준의 절대오차(MAE, RMSE)와 집단 간 상관(Pearson r , Spearman ρ)을 산출

하였으며, 결과는 <표 4.3>과 같다.

분석 결과를 살펴보면 어느 지표를 기준으로 삼느냐에 따라 모델의 재현 성능 평가가 달라졌다. 응답률의 절대적 근접성(MAE)과 집단 간 순위 구조의 재현(상관)이 모델 별로 서로 다르게 나타났기 때문이다. Gemini 2.5 Pro는 Q1에서 집단 수준 오차가 가장 작았으나(MAE .083) 집단 간 상관은 중간 수준($r=.441$)에 머물렀다. 반면 Claude Sonnet 4.5는 절대 수준은 실제보다 높으면서도 집단 간 순위 구조의 재현은 상대적으로 우수하였다(Q2 $r=.786$). Gemini 2.5 Flash는 집단 간 상관이 사실상 0에 가까워($r=.009\sim.041$) 실제 인구집단의 응답 구조를 거의 재현하지 못하였다. Claude Haiku 4.5는 실제 응답률과의 차이가 가장 컸으나(MAE .503, .647), 집단 간 상관은 Gemini Flash보다 높았다. 결국 네 모델 중 응답 수준과 집단 간 패턴을 모두 안정적으로 재현한 모델은 없었으며, 어느 지표에서 우수한지도 모델마다 달랐다. 이는 동일한 페르소나를 입력하더라도 재현 성능이 모델에 따라 달라짐을 보여준다.

이러한 모델 간 차이는 모델 특성과도 어느 정도 연관되어 있을 것으로 추측된다. <표 4.3>에서 계열 내 비교는 일관되었다. Gemini 2.5 Pro는 Gemini 2.5 Flash보다, Claude Sonnet 4.5는 Claude Haiku 4.5보다 두 문항 모두에서 집단 수준 절대오차(MAE)가 작았다. 반면 계열 간 비교는 문항에 따라 달랐다. Q1에서는 Claude Sonnet 4.5(.181)가 Gemini 2.5 Flash(.167)보다 오차가 컸다. Q2에서는 Claude Sonnet 4.5(.238)가 네 모델 중 가장 작은 오차를 보여 Gemini 2.5 Pro(.273)를 앞섰다. 계열 평균에서 관찰되는 Gemini 계열의 우위는 상당 부분 Claude Haiku 4.5의 큰 오차에 기인한다. 이러한 결과는 LLM 모델의 성능과 개발사에 따라 응답 경향이 어떻게 나타나는지를 탐색적으로 살펴본 것으로, 그 결과를 모델 규모나 계열 전반의 특성으로 일반화할 수 없다.

<표 4.3> 하위집단 수준 분포 재현성 지표

모델	문항	MAE	RMSE	평균편향	Pearson r	Spearman ρ
Gemini 2.5 Pro	Q1	.083	.100	+.023	.441	.384
	Q2	.273	.285	+.273	.454	.356
Gemini 2.5 Flash	Q1	.167	.214	+.158	.041	-.030
	Q2	.355	.368	+.355	.009	-.042
Claude Sonnet 4.5	Q1	.181	.193	+.181	.680	.698
	Q2	.238	.258	+.238	.786	.596
Claude Haiku 4.5	Q1	.503	.512	+.503	.220	.192
	Q2	.647	.650	+.647	.496	.497

주: 성별×연령×지역 18개 집단별 합성 응답률과 실제 응답률(가중)을 대응시켜 산출. MAE·RMSE·평균편향은 집단 수준 절대 오차, $r\cdot\rho$ 는 집단 간 응답률 수준·순위 구조의 상관

앞서 4.1절에서 확인된 데이터셋의 학력·주거형태·혼인상태 분포 편차가 관찰된 간극의 원인일 가능성을 검토하였다. 세 변수의 분포가 기부 의향과 상관될 경우, 간극이 모델 특성이 아니라 데이터셋의 구성 편향에서 비롯되었을 수 있기 때문이다. 이를

확인하기 위해 성별×연령×지역 18개 집단의 결합 분포를 사후충화로 맞춘 상태에서, 학력·주거형태·혼인상태는 각각의 주변 분포만 사회조사(2025) 기준에 맞추는 반복 비례가중(raking) 민감도 분석을 수행하였다(<표 4.4>). 그 결과 보정에 따른 응답률 변화는 모든 모델과 문항에서 +1.1~+4.0%p로 나타났으며, 그 방향은 모두 양(+)으로 나타나 보정 후 간극은 Q1에서 최대 +52.6%p, Q2에서 최대 +67.3%p로 오히려 소폭 확대되었다. 즉 관찰된 과대추정이 데이터셋의 분포 편향에서 비롯되었을 가능성은 낮으며, 오히려 모델 수준의 응답 특성과 관련되어 있을 가능성이 있다. 다만 이 해석은 재가중 분석 대상인 세 변수에 한정되며, 그 외 특성의 영향까지 배제한 것은 아니다.

<표 4.4> 학력·주거형태·혼인상태 반복비례가중 민감도 분석

모델	문항	무보정	보정 후	보정효과	보정 후 간극
Gemini 2.5 Pro	Q1	43.9%	46.4%	+2.5%p	+6.3%p
	Q2	50.4%	53.2%	+2.8%p	+31.5%p
Gemini 2.5 Flash	Q1	58.3%	60.9%	+2.6%p	+20.8%p
	Q2	59.2%	62.1%	+2.9%p	+40.4%p
Claude Sonnet 4.5	Q1	58.1%	61.6%	+3.5%p	+21.5%p
	Q2	45.6%	49.6%	+4.0%p	+27.9%p
Claude Haiku 4.5	Q1	91.6%	92.7%	+1.1%p	+52.6%p
	Q2	87.6%	89.0%	+1.4%p	+67.3%p

주1: 보정효과와 간극은 표시값 기준으로 산출하였으며, 반복비례가중 적용 후 유효표본크기(ESS)는 1,285(무보정 1,447)이다.

주2: 혼인상태의 별거 범주는 Nemotron-Personas-Korea에 존재하지 않아 법적 혼인이 유지되는 상태임을 고려하여 배우자있음에 포함하였다.

LLM의 합성 응답과 사회조사 응답의 차이를 분석한 결과 모든 모델에서 실제 사회조사보다 동의율이 높게 나타났다. 사회조사에서 뚜렷했던 기부 의향(Q1)과 유산 기부 의향(Q2) 간 차이를 합성 응답은 재현하지 못하였고, 결과적으로 합성-실제 응답의 간극은 Q2에서 더 크게 나타났다. 또한 집단 간 응답 구조의 재현 정도에서도 모델 간 차이가 뚜렷하여, 페르소나 기반 응답 시뮬레이션이 실제 인구의 응답 분포를 그대로 재현하기보다 모델 고유의 응답 성향을 상당 부분 반영하고 있는 것으로 보인다.

4.3 프레이밍 실험 결과

5개 문항 버전(원문, Q2_A~Q2_D)에 대한 4개 모델의 응답을 비교한 결과, 문항 프레이밍은 LLM의 응답 분포에 영향을 미치는 것으로 나타났다. 4개 모델의 원문 평균 동의율은 60.1%였으나, 네 프레이밍 문항 모두에서 동의율이 감소하였다. 이는 H1~H3의 예측과는 부합하였으나, 동의율 상승을 예측한 H4는 부합하지 않았다. 평균 감소

폭은 구체적 행동 의도를 묻는 Q2_B가 가장 컸으며(-55.3%p), 다음으로 Q2_A(조건부 표현 제거, -29.4%p), Q2_C(손실 프레임, -26.4%p), Q2_D(상징적 행위, -23.1%p) 순이었다.

프레이밍 효과의 검정은 동일 페르소나가 원문과 프레이밍 문항에 반복 응답한 대응 구조를 반영하여 정확 McNemar 검정으로 수행하였다(<표 4.5>). 모든 모델과 프레이밍 문항 조합에서 원문 대비 동의율이 유의하게 감소하였으며($p < .001$), 신뢰구간은 전 구간에서 0을 포함하지 않았다. 특히 응답 변화의 방향이 뚜렷하게 비대칭적이었다. 생성의 무작위성에 의한 번복이라면 '있다→없다'와 '없다→있다'가 유사한 빈도로 나타나야 하지만, 관찰된 번복은 '있다→없다' 방향에 집중되었다(예: Claude Haiku의 Q2_B에서 207건 대 0건). 이는 번복이 확률적 잡음이 아니라 문항 표현 변화에 대한 체계적 반응일 가능성을 보여준다.

가장 두드러진 결과는 Q2_B에서 나타났다. 구체적 행동 의도를 직접 묻는 방식으로 문항을 수정하자 동의율이 대부분의 모델에서 매우 낮게 나타났다. Claude Sonnet 4.5 (0.4%)와 Gemini 2.5 Pro(0.7%)에서는 사실상 긍정 응답이 거의 나타나지 않았으며, Claude Haiku 4.5(8.5%)와 Gemini 2.5 Flash(9.6%) 역시 실제 사회조사 결과(21.7%)를 하회하였다. 이는 추상적 가치 지향 질문과 구체적 행동 의도를 요구하는 질문을 LLM이 다르게 처리할 수 있음을 보여준다.

Q2_D의 결과는 이론적 예측과 달리 전 모델에서 원문 대비 동의율 감소(-13.3~ -34.1%p)로 나타나 H4는 지지되지 않았다. 이는 2.3절에서 논의한 바와 같이, 상징성 강조에 따른 규범적 신호의 상방 효과보다 유연장 작성이라는 행위의 구체성 기제가 우세하게 작동했을 가능성이 있다. 다만 Q2_D의 동의율(12.6~71.9%)은 모든 모델에서 순수하게 구체적 계획을 묻는 Q2_B(0.4~9.6%)를 뚜렷하게 상회하여 상징적 의미 부여가 구체성에 따른 동의율 하락을 일부 완충했을 가능성도 있다.

<표 4.5> 원문 대비 프레이밍 문항의 응답 변화 (McNemar 검정)

모델	문항	원문 동의율	프레이밍 동의율	차이 [95% CI]	있다→ 없다	없다→ 있다
Gemini 2.5 Pro	Q2_A	0.504	0.215	-.289 [-.344, -.234]	79	1
	Q2_B	0.504	0.007	-.496 [-.556, -.437]	134	0
	Q2_C	0.504	0.363	-.141 [-.195, -.086]	50	12
	Q2_D	0.504	0.322	-.181 [-.237, -.126]	58	9
Gemini 2.5 Flash	Q2_A	0.581	0.337	-.244 [-.299, -.190]	69	3
	Q2_B	0.581	0.096	-.485 [-.545, -.426]	131	0
	Q2_C	0.581	0.322	-.259 [-.321, -.198]	80	10
	Q2_D	0.581	0.311	-.270 [-.326, -.214]	76	3
Claude Sonnet 4.5	Q2_A	0.467	0.059	-.407 [-.466, -.349]	110	0
	Q2_B	0.467	0.004	-.463 [-.522, -.403]	125	0
	Q2_C	0.467	0.215	-.252 [-.305, -.199]	69	1
	Q2_D	0.467	0.126	-.341 [-.397, -.284]	92	0

<표 4.5> 원문 대비 프레이밍 문항의 응답 변화 (McNemar 검정)(계속)

모델	문항	원문 동의율	프레이밍 동의율	차이 [95% CI]	있다→ 없다	없다→ 있다
Claude Haiku 4.5	Q2_A	0.852	0.615	-.237 [-.291, -.183]	67	3
	Q2_B	0.852	0.085	-.767 [-.817, -.716]	207	0
	Q2_C	0.852	0.448	-.404 [-.463, -.344]	110	1
	Q2_D	0.852	0.719	-.133 [-.176, -.090]	38	2

주: 모든 조합 $p < .001$

나아가 생성 확률성의 영향을 정량적으로 분리하기 위해 관찰 번복률에서 각 모델과 문항의 자기 불일치율(1-자기 일치율)을 차감한 초과 번복률을 산출하였다(<표 4.6>). 16개 조합 모두에서 초과 번복률이 양(+)의 값을 보여, 프레이밍에 따른 응답 변화가 생성 확률성만으로는 설명되지 않음을 확인하였다. 초과 번복률은 구체적 계획을 묻는 Q2_B에서 가장 컸다(모델별 .394~.754). 다만 기저선 보정의 효과는 모델에 따라 달랐는데, 자기 일치율이 높은 Claude 계열은 관찰 번복률이 거의 그대로 초과 번복률로 이어진 반면, Gemini Flash는 관찰 번복률의 상당 부분이 생성 확률성에서 기인하여 보정 후 크기가 절반 이하로 줄어드는 조합도 있었다(Q2_A: .267 → .121). 이는 반복 검증 없이 단일 생성 응답만으로 번복률을 해석하는 것의 위험을 보여준다.

프레이밍 효과는 응답 수준을 일부 조정하는 수준이 아니라 응답 분포 자체를 변화시키는 것으로 보인다. 특히 구체적 행동 의도를 요구하는 문항(Q2_B)은 가장 큰 동의율 감소와 가장 높은 초과 번복률을 동시에 보였다. 또한 Q2_A, Q2_C, Q2_D 등 일부 프레이밍은 원문 대비 동의율을 낮추어 실제 사회조사 결과에 근접한 응답 분포를 생성 하기도 하였다. 이러한 결과는 LLM 기반 사회 시뮬레이션의 결과가 모델 자체뿐 아니라 문항 설계 방식에도 크게 좌우될 수 있음을 보여준다. 다만 각 프레이밍이 어떤 인지 기제를 통해 작동하는지는 문항별로 여러 조작이 겹쳐 있어, 본 연구의 가설 관련 논의는 인과적 검증이 아니라 탐색적 관찰로 제한한다.

<표 4.6> 자기 불일치율 기저선 대비 초과 번복률

모델	Q2_A	Q2_B	Q2_C	Q2_D
Gemini 2.5 Pro	0.198	0.481	0.114	0.127
Gemini 2.5 Flash	0.121	0.394	0.198	0.122
Claude Sonnet 4.5	0.400	0.463	0.249	0.336
Claude Haiku 4.5	0.205	0.754	0.352	0.109

4.4 모델 간 응답 일관성 분석

앞선 분석에서는 각 모델의 응답 분포를 실제 사회조사 결과와 비교하였다. 그러나 실리온 샘플링의 활용 가능성을 평가하기 위해서는 동일한 페르소나에 대해 서로 다른 모델이 얼마나 일관된 응답을 생성하는지도 검토할 필요가 있다.

다만 LLM 응답은 동일한 입력에 대해서도 실행마다 달라질 수 있는 확률적 산출물이므로, 모델 간 일치율을 해석하기에 앞서 동일 모델을 반복 실행했을 때의 자기 일치율(self-consistency)을 기저선으로 확인하였다. <표 4.7>은 프레이밍 실험 표본 270개에 대해 6개 문항을 모델별로 3회 반복 생성하여(3.4절 참조) 산출한 쌍별 자기 일치율이다. 자기 일치율은 전체 평균 기준 85.1%(Gemini 2.5 Flash)에서 99.2%(Claude Sonnet 4.5)에 이르러, 네 모델 모두 자신의 응답을 대체로 안정적으로 재생산하였다.

<표 4.7> 모델별 자기 일치율 (270명 × 6문항 × 3회 반복)

구분	Q1	Q2	Q2_A	Q2_B	Q2_C	Q2_D	전체 평균
Gemini 2.5 Pro	.859	.864	.901	.985	.884	.879	.895
Gemini 2.5 Flash	.830	.817	.854	.909	.864	.830	.851
Claude Sonnet 4.5	.983	.993	.993	1.000	.990	.995	.992
Claude Haiku 4.5	.975	.965	.946	.988	.941	.960	.963

이 기저선 위에서 모델 간 일치율을 산출하였다(<표 4.8>). 생성 확률성에 의한 불일치를 최소화하기 위해 각 모델의 3회 반복 다수결 응답을 비교 기준으로 사용하였다. 6개 모델 쌍의 전체 일치율은 62.4%(Claude Sonnet × Claude Haiku)에서 85.6%(Gemini Pro × Gemini Flash) 범위로, 모든 쌍에서 해당 두 모델 자기 일치율의 평균을 하회하였다. 특히 Claude 계열이 포함된 쌍에서는 자기 일치율(96.3~99.2%)과의 격차가 컸던 반면, Gemini 두 모델 간 일치율(85.6%)은 Gemini Flash의 자기 일치율(85.1%)과 유사한 수준이었다. 다수결 기준에서 생성 확률성이 상당 부분 제거된 점을 고려하면, 관찰된 모델 간 불일치(14.4~37.6%)는 무작위 변동이 아니라 페르소나 정보를 응답으로 변환하는 방식의 모델 간 체계적 차이로 보인다.

<표 4.8> 모델 간 응답 일치율 (다수결 응답 기준)

구분	Q1	Q2	Q2_A	Q2_B	Q2_C	Q2_D	전체 평균
Gemini 2.5 Pro × Gemini 2.5 Flash	.811	.789	.822	.948	.885	.881	.856
Gemini 2.5 Pro × Claude Sonnet 4.5	.763	.804	.848	.993	.781	.774	.827
Gemini 2.5 Flash × Claude Sonnet 4.5	.811	.726	.707	.941	.807	.774	.794
Gemini 2.5 Flash × Claude Haiku 4.5	.659	.719	.652	.952	.733	.585	.717
Gemini 2.5 Pro × Claude Haiku 4.5	.537	.656	.570	.930	.789	.593	.679
Claude Sonnet × Claude Haiku 4.5	.678	.593	.448	.930	.689	.404	.624

주목할 것은 일치율의 구조가 개발사 단위로 묶이지 않았다는 점이다. 가장 낮은 일치율은 오히려 동일 개발사의 Claude Sonnet × Claude Haiku 쌍(62.4%)에서 나타났으며, 특히 Q2_D(40.4%)와 Q2_A(44.8%)에서 절반 이하로 떨어졌다. 이는 Claude Haiku의 전반적으로 높은 동의율 경향이 동일 개발사 모델과의 일치조차 낮추었기 때문으로, 합성 응답의 특성이 모델 계열이 아니라 개별 모델 수준에서 결정됨을 보여준다. 따라서 본 연구의 모델 간 비교는 특정 계열의 우열이 아니라 본 연구에 포함된 4개 모델 내의 탐색적 비교로만 해석되어야 한다.

문항별로는 Q2_B의 모델 간 일치율이 전 쌍에서 93.0~99.3%로 예외적으로 높았는데, 이는 모델 간 수렴이라기보다 모든 모델의 동의율이 0.4~9.6%로 바닥에 근접한 데 따른 결과이므로, 높은 일치율 자체를 재현성의 근거로 해석하는 것은 적절하지 않다. 반대로 Q1이나 Q2 원문처럼 모델의 동의율이 중간 수준에 분포하는 문항에서는 자기 일치율과 모델 간 일치율이 모두 상대적으로 낮게 나타났다. 이는 응답 확률이 0.5에 가까울수록 생성 확률성과 모델 간 판단 차이가 관찰될 여지가 커지기 때문으로 해석된다.

끝으로 반복 검증이 본실험으로부터 일정 기간 경과 후 수행된 점을 이용하여 시점 간 안정성을 확인하였다. 원실험 응답과 반복 생성 응답 간 일치율은 모든 모델과 문항 조합에서 동시점 자기 일치율과 $\pm 3.7\%p$ 이내의 차이를 보여, 관찰된 응답 분포가 호출 시점이나 클라이언트 환경 변화에 민감하지 않은 것으로 나타났다(부록 참조).

실리콘 샘플링은 모델 간 일정 수준의 공통성을 보이면서도 동시에 상당한 수준의 모델 의존성을 내포하고 있다. 다수결로 생성 확률성을 제거한 뒤에도 모델 간 불일치가 최대 37.6%에 이르렀고, 그 구조가 개발사 단위로 묶이지 않았다는 점은 합성 응답이 개별 모델 수준에서 달라질 수도 있음을 보여준다. 따라서 LLM 기반 사회조사를 활용할 때에는 단일 모델 결과에 의존하기보다 다중 모델 검증을 병행할 필요가 있을 것이다.

5. 결론 및 함의

5.1 연구 결과 요약

본 연구는 Nemotron-Personas-Korea를 활용하여 한국형 실리콘 샘플링의 가능성과 한계를 검토하였다. 주요 결과는 다음과 같다.

첫째, Nemotron-Personas-Korea 데이터셋은 성별·연령·지역 분포에서 사회조사와 상당히 유사하였으나, 학력·주거형태·혼인상태 분포에서는 상대적으로 편차가 높은 것을 확인하였다. 이는 합성 페르소나의 재현 수준이 변수의 성격에 따라 달라질 수 있음을 시사하며, 활용 목적에 따라 어떤 변수의 분포가 중요한지를 함께 검토할 필요가 있음을 보여준다.

둘째, 직업·가족관계·문화적 배경이 포함된 다차원 한국형 페르소나를 활용하였음에도 모든 모델에서 실제 사회조사보다 높은 기부 의향이 관찰되었다. 가중 기준 간극은 Q1에서 $+3.8 \sim +51.5\%p$, Q2에서 $+23.9 \sim +65.9\%p$ 였으며, 효과크기는 작음에서 큼(Cohen's

$h=0.08\sim1.45$)에 이르는 범위로 나타났다. 특히 학력·주거형태·혼인상태 분포를 사회조사 기준으로 재가중한 반복비례가중 분석에서도 응답률 변화가 $+1.1\sim+4.0\%$ p에 그치고 그 방향이 모두 양(+)으로 나타나 간극이 오히려 소폭 확대되어, 관찰된 간극이 데이터셋의 분포 편향만으로 설명되지 않음을 확인하였다.

셋째, 프레이밍 실험 결과 문항 표현 방식은 응답 분포에 큰 영향을 미쳤다. 네 프레이밍 문항 모두 원문 대비 동의율을 낮추었다. 특히 구체적인 행동 의도를 묻는 문항(Q2_B)은 평균 55.3%p의 감소 효과를 보였다. 반면 상징성 강조에 따른 동의율 상승을 예측한 H4는 지지되지 않았는데, 이는 규범적 신호보다 행위 구체성이 LLM 응답에 더 강하게 작용할 수 있음을 보여주는 결과이다. 자기 불일치율을 차감한 초과 번복률 역시 16개 조합 전부에서 양수로 나타나, 이러한 변화가 생성 확률성만으로는 설명되지 않음을 확인하였다.

넷째, 동일한 페르소나에 대해서도 모델 간 응답 차이가 상당하였다. 생성 확률성을 제거한 다수결 기준에서도 모델 간 불일치가 최대 37.6%에 이르렀고, 그 구조가 개발사 단위로 묶이지 않아 합성 응답이 개별 모델의 응답 성향에 좌우될 수 있음을 보여주었다.

5.2 이론적 기여

본 연구는 LLM 기반 실리콘 샘플링 연구에 두 가지 함의를 제공한다. 첫째, 페르소나의 현실성이 향상되더라도 모델 고유의 응답 성향이 지속되는 패턴을 관찰하였다. 본 연구는 다차원 페르소나를 활용하여 페르소나 정교화 가설과 구조적 편향 가설을 대비하고자 하였으나, 연구 설계상 두 가설 중 하나를 확정적으로 지지하거나 기각하기는 어렵다. 다만 정교한 페르소나에도 간극이 유지되고, 그 간극이 데이터셋의 분포 편향을 보정한 뒤에도 거의 그대로 남았다는 점은 간극의 상당 부분이 페르소나 정보의 부족보다 모델 수준의 응답 특성과 관련되어 있을 가능성을 시사한다. 이는 실리콘 샘플링의 정확도가 페르소나 정교성만으로 결정되지 않음을 보여주는 관찰적 근거로 이해될 수 있다.

둘째, 프레이밍 효과가 인간 응답뿐 아니라 LLM 응답에서도 강하게 작동함을 확인하였다. 행동의 구체성, 조건부 표현, 손실 프레임과 같은 언어적 요소는 응답 분포를 크게 변화시켰다. 다만 각 프레이밍 문항은 하나의 인지적 요소만을 분리하지 못하며, 예컨대 Q2_B는 행동 구체성과 함께 시제·확정성이, Q2_C는 손실 프레임과 함께 준거집단 대비가, Q2_D는 상징성과 함께 행위 구체성이 동시에 변화한다. 따라서 본 연구의 프레이밍 결과는 개별 인지 기제에 대한 인과적 검증이 아니라, 문항 설계가 LLM 응답에 미치는 영향에 관한 탐색적 관찰로 해석되어야 한다. 그럼에도 문항 구성 자체가 LLM 기반 조사 설계에서 중요한 방법론적 변수임을 보여준다.

5.3 행정 및 정책 연구자를 위한 가이드라인

본 연구 결과는 실리콘 샘플링을 행정 및 정책 연구에 활용할 때 다음과 같은 점을 고려할 필요가 있음을 보여준다.

첫째, 실리콘 샘플링은 정책 수용성이나 사회 인식 변화에 대한 초기 탐색, 가설 생성, 시나리오 비교 분석에는 유용하게 활용될 수 있다.

둘째, 특정 응답물의 절댓값을 예측하는 용도로는 신중한 접근이 필요하다. 본 연구에서 모든 모델이 실제 조사보다 높은 응답률을 보인 것처럼, LLM 기반 응답은 체계적 편향을 포함할 가능성이 있다.

셋째, 문항 설계가 결과에 큰 영향을 미치므로 단일 문항만으로 결론을 도출하기보다 다양한 프레이밍을 병행 검토할 필요가 있다. 특히 구체적인 행동 의도나 손실 프레임은 응답 편향을 완화하는 유용한 방법이 될 수 있을 것으로 예상된다.

넷째, 모델 간 차이가 크게 나타났으므로 단일 모델에 의존하기보다 다중 모델 검증을 수행하는 것이 바람직하다.

5.4 한계 및 후속 연구

본 연구는 기부 의향과 유산 기부 의향이라는 특정 주제를 중심으로 수행되었기 때문에 결과를 모든 정책 영역으로 일반화하는 데에는 한계가 있다. 또한 Google 및 Anthropic 계열 모델만을 대상으로 하였으며, 한국어 특화 LLM이나 국내 개발 모델과의 비교는 수행하지 못하였다.

방법론적으로도 몇 가지 제약이 있다. 첫째, 응답 생성은 temperature 1.0의 확률적 설정에서 이루어졌다. 이는 개별 페르소나의 결정론적 응답이 아니라 집단 수준의 응답 분포를 재현하기 위한 의도적 선택이었으나, 그 결과 개별 응답 쌍을 비교하는 지표에는 생성 확률성이 개입한다. 본 연구는 이를 프레이밍 표본 270명에 대한 반복 검증으로 정량화하였으나, 전체 표본 1,800명에 대한 반복 검증은 비용 제약으로 수행하지 못하였다.

둘째, 모델 버전의 식별 가능성이 제공사에 따라 달랐다. Claude 계열은 날짜 고정 스냅샷으로 호출하여 본실험과 반복 검증 시점 간 모델 동일성을 확보하였으나, Gemini 2.5 Pro·Flash는 Vertex AI가 단일 안정 모델 계열로만 제공하고 날짜 고정 버전을 지원하지 않아, 두 시점에 동일한 하위 빌드가 사용되었음을 직접 증명할 수는 없다. 다만 두 시점의 응답 자기 일치율이 $\pm 3.7\%p$ 이내로 유지되어(부록3), 잠재적 버전 갱신이나 클라이언트 환경 변화가 관찰된 응답 분포에 미친 영향은 제한적이었던 것으로 판단된다.

셋째, 본 연구에서 비교한 고성능·경량 및 개발사 계열 구분은 내부 추론 활성화 여부 등 다른 요인과 교란되어 있어, 그 차이를 특정 요인으로 귀속하기는 어렵다.

후속 연구에서는 복지 수용성, 정책 신뢰, 출산, 이민, 주거 정책 등 다양한 정책 분야로 분석 범위를 확대할 필요가 있다. 또한 전체 표본에 대한 반복 샘플링을 통한 안정성 검증, 한국어 특화 LLM과의 비교, 프레이밍 효과의 재현성 검증 등이 추가로 이루어질 필요가 있다. 특히 재현성 확보를 위해서는 날짜 고정 스냅샷을 제공하는 모델을 우선 선정하거나, 버전 고정이 불가능한 경우 응답 수집 시점의 버전 메타데이터를 함께 기록하는 절차를 마련할 필요가 있다. 나아가 실리콘 샘플링을 정책 시뮬레이션이나 사회 시스템 모델링과 결합하는 활용 가능성 역시 중요한 연구 과제이다.

참고문헌

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.
- Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023, September 22). *Which Humans?*. <https://doi.org/10.31234/osf.io/5b26t>
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416.
- Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine bias: How do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3), 1156–1196.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). *Scaling synthetic data creation with 1,000,000,000 personas*. arXiv preprint arXiv:2406.20094.
- Horton, J. J., Filippas, A., & Manning, B. S. (2026). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. 31122, rev. Feb. 2026). National Bureau of Economic Research. <https://doi.org/10.3386/w31122>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.
- Lee, S., Yang, K. Q., Peng, T. Q., Heo, R., & Liu, H. (2024). Exploring social desirability response bias in large language models: Evidence from GPT-4 simulations. arXiv preprint arXiv:2410.15442.
- Ma, B., Yoztyurk, B., Haensch, A. C., Wang, X., Herklotz, M., Kreuter, F., Plank, B., & Aßenmacher, M. (2025). Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1785–1809). Association for Computational Linguistics.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). *Generative agent simulations of 1,000 people*. arXiv preprint arXiv:2411.10109.
- Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12), pgae533.

- Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage Publications.
- Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024). *Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information*. arXiv preprint arXiv:2402.18144.
- Tjautja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics*, 12, 1011–1026.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.
- Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., & Su, W. J. (2024). *On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization*. arXiv preprint arXiv:2405.16455.

부록

부록1. 응답 생성 설정

본 연구의 모든 응답은 아래 설정에 따라 생성되었다. 본실험(1,800개)과 반복 검증(270개) 모두 동일한 프롬프트와 응답 코딩 규칙을 적용하였으며, 생성 설정 또한 동일하다.

항목	내용
모델 및 버전	Claude 계열은 날짜 고정 스냅샷으로 지정하여 시점 간 동일성을 확보하였다: Claude Sonnet 4.5(claude-sonnet-4-5-20250929), Claude Haiku 4.5(claude-haiku-4-5-20251001). Gemini 2.5 Pro·Flash는 Google Cloud가 단일 안정 모델 ID(2025년 6월 17일 출시, 하위 스냅샷 미제공)로 관리하므로 안정 모델 계열 단위로 통제하였다.
디코딩 설정	temperature 1.0, top-p 기본값 페르소나 집단의 응답 분포 재현을 위해 각 응답을 모델 분포로부터의 무작위 추출로 간주할 수 있는 temperature 1.0을 채택하였다.
내부 추론(Thinking)	Gemini 2.5 Pro·Flash는 Thinking이 활성화된 상태로 호출하였으며 (3.2절의 계열 간 대비 설계), 사고 토큰 및 청구 내역에서 전 구간 작동을 확인하였다. Claude 계열은 별도 추론 과정을 수행하지 않는다.
생성 횟수	본실험: 페르소나·문항 조합당 1회 반복 검증: 270개 × 6문항(Q1, Q2, Q2_A, Q2_B, Q2_C, Q2_D) × 4모델 × 3회(총 19,440회)
본실험 호출 시점	2026년 5월 17일 2026년 5월 22일
반복 검증 호출 시점	2026년 7월 12일
SDK(Software Development Kit)	Claude는 두 시점 모두 anthropic SDK로 통일. Gemini는 본실험 vertexai.generative_models, 반복검증 google-genai를 사용하였으며, SDK 변화의 영향까지 포함해 시점 간 안정성을 확인하였다(부록3).
응답 형식	각 문항에 대한 이분형 응답(예/아니오)

부록2. 프롬프트 템플릿 및 응답 코딩 규칙

각 페르소나에 대해 아래 형식의 프롬프트를 구성하여 응답을 생성하였다. {페르소나 서술}, {professional_persona}, {family_persona}, {cultural_background}에는 Nemotron-Personas-Korea의 서술형 페르소나가, {문항}에는 해당 검증 문항 또는 프레이밍 문항이 삽입된다.

[프롬프트 템플릿]

다음은 한 사람에 대한 소개입니다.

{페르소나 서술}

직업: {professional_persona}

가족: {family_persona}

문화적 배경: {cultural_background}

위 인물의 입장에서 다음 질문에 답해주세요.

반드시 보기 중 하나만 선택하고, 선택한 번호만 답하세요.

질문: {문항}

1. 있다

2. 없다

답변 (번호만):

[응답 코딩 규칙]

모델의 자연어 응답에서 최초로 등장하는 1 또는 2를 추출하여 원자료로 저장하였다 (1=있다, 2=없다). 응답률 및 일치율 산출 시 1(있다) 응답을 동의로 간주하여 그 비율을 집계하였다.

부록3. 시점 간 일치율과 동시점 자기 일치율 비교

반복검증(2026년 7월 12일)이 본실험(2026년 5월)으로부터 약 2개월 경과 후 수행된 점을 이용하여, 원실험 응답과 반복 생성 응답 간의 시점 간 일치율을 산출하고 이를 동시점 자기 일치율과 비교하였다. 모든 모델과 문항 조합에서 두 값의 차이는 $\pm 3.7\%$ 이내였다. 이는 각 조합의 동시점 자기 불일치율(0.0~18.3%)과 비교할 때 대체로 그 범위 안에 드는 수준으로, 시점 간 차이가 반복 생성에 내재한 확률적 변동을 크게 벗어나지 않음을 뒷받침한다. 아울러 SDK(Software Development Kit)가 두 시점 모두 동일했던 Claude 계열에서도 최대 2.5%p의 시점 간 차이가 관찰되어, 관찰된 차이가 Gemini의 SDK 변경 자체보다는 생성 확률성에서 주로 기인했을 가능성을 보여준다. 다만 Gemini 계열은 날짜 고정 스냅샷 확인이 어려워 두 시점의 모델 버전 동일성을 직접 확인할 수는 없으므로(5.4절 참조), 호출 시점과 환경 변화가 응답 분포에 미친 영향이 있었더라도 그 크기는 제한적이었던 것으로 판단된다.

모델	문항	시점 간 일치율	동시점 자기 일치율	차이
Gemini 2.5 Pro	Q1	0.844	0.859	0.015
	Q2	0.827	0.864	0.037
	Q2_A	0.895	0.901	0.006
	Q2_B	0.984	0.985	0.001
	Q2_C	0.886	0.884	-0.002
	Q2_D	0.872	0.879	0.007
Gemini 2.5 Flash	Q1	0.828	0.830	0.002
	Q2	0.833	0.817	-0.016
	Q2_A	0.860	0.854	-0.006
	Q2_B	0.916	0.909	-0.007
	Q2_C	0.847	0.864	0.017
	Q2_D	0.841	0.830	-0.011
Claude Sonnet 4.5	Q1	0.968	0.983	0.015
	Q2	0.969	0.993	0.024
	Q2_A	0.981	0.993	0.012
	Q2_B	1.000	1.000	0.000
	Q2_C	0.965	0.990	0.025
	Q2_D	0.986	0.995	0.009
Claude Haiku 4.5	Q1	0.973	0.975	0.002
	Q2	0.953	0.965	0.012
	Q2_A	0.947	0.946	-0.001
	Q2_B	0.985	0.988	0.003
	Q2_C	0.947	0.941	-0.006
	Q2_D	0.960	0.960	0.000

분석용 데이터

1. 합성 페르소나 데이터

Kim, H., Ryu, J., Lee, J., Ryu, H., Praveen, K., Prayaga, S., Thadaka, K., Jennings, W., Sadeghi, B., Sharabiani, A., Choi, Y., & Meyer, Y. (2026). *Nemotron-Personas-Korea: Synthetic personas aligned to real-world distributions for Korea* [Dataset]. Hugging Face. <https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea>

2. 사회조사 데이터

국가데이터처. (2025). 사회조사 [마이크로데이터]. 국가데이터처 마이크로데이터 통합 서비스(MDIS). <https://doi.org/10.23333/PN.20006531.V2.2>

Can Korean Synthetic Personas Reproduce Real Survey Responses?: The Possibilities and Limitations of Silicon Sampling Using the Nemotron-Personas-Korea Dataset

Gwangsu Eo²⁾

Abstract

This study examines the potential and limitations of LLM-based Silicon Sampling using Nemotron-Personas-Korea. Responses generated by four LLMs (Gemini and Claude series) for 1,800 synthetic Korean personas were compared with the 2025 Korean Social Survey. In addition, framing experiments were conducted on a subsample of 270 personas using four alternative question formats. The results show that all models overestimated charitable and bequest donation intentions compared with the survey results, with larger discrepancies observed for bequest donation. Question framing substantially influenced response distributions, and questions requiring concrete behavioral commitment produced the largest reduction in agreement rates. Considerable variation was also observed across models for the same personas. These findings suggest that refined personas alone cannot fully eliminate the gap between synthetic and real responses, while questionnaire design plays a critical role in shaping LLM-generated survey outcomes.

Key words : Silicon Sampling, Large Language Models, Synthetic Personas, Framing Effects, Social Survey

2) Single Author, Research Fellow, Population Impact Assessment Center, Gyeonggi Research Institute, E-mail: egs@gri.re.kr