

Synthetic Population

Nemotron-Personas와 한국 synthetic population

: 인구통계 기반으로 가상 인물을 생성하고,

각 가상 인물에게 성격과 상황별 페르소나 설명을 붙인 합성 데이터

IDEA Lab 안영민

**Korea Extended
2026**

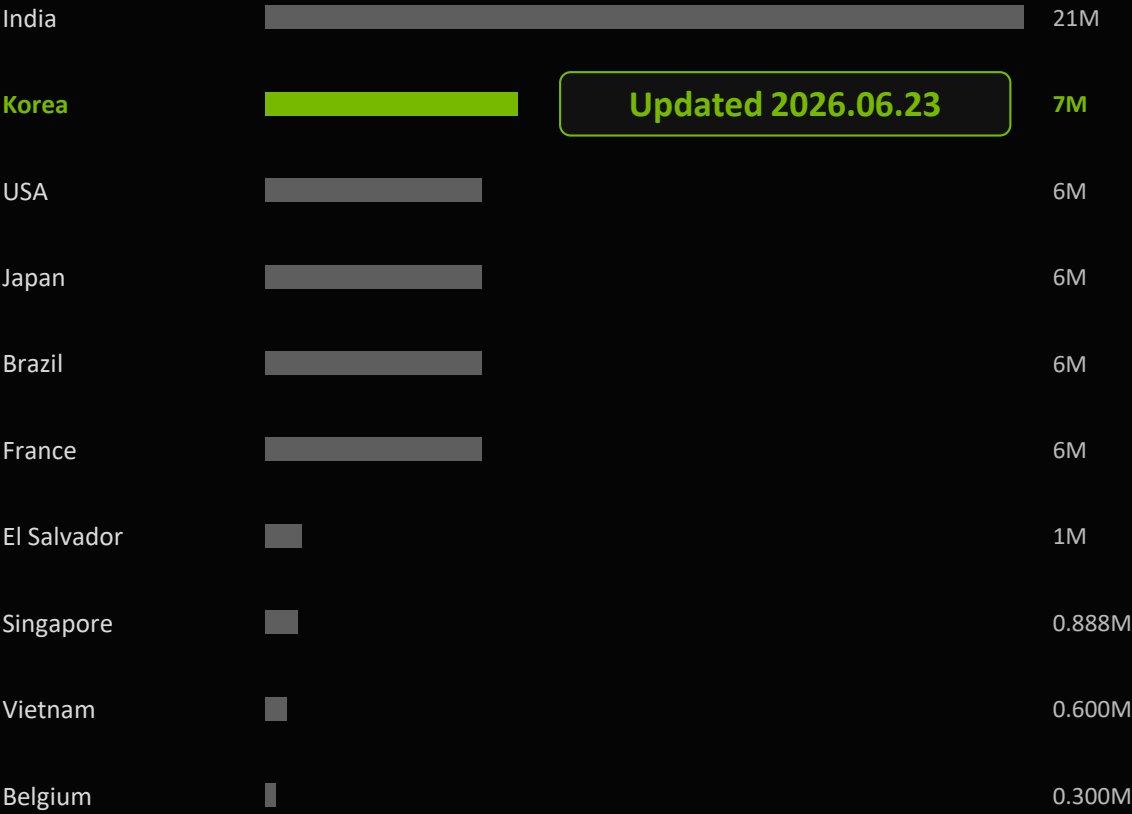
**PGM + OCEAN + LLM
분포 기반 합성 페르소나**

1. 지금까지 어떤 synthetic population을 만들었나



공식 dev note의 지도

페르소나 설명 수



2. 한국 synthetic population은 어떻게 생겼나

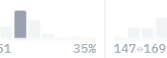
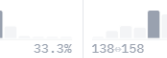
[Nemotron-Personas-Korea](#): 대한민국 분포 기반 합성 페르소나 데이터셋

1M
레코드

26
변수

7
상황별 페르소나

- 페르소나 변수

professional_persona string · lengths	sports_persona string · lengths	arts_persona string · lengths	travel_persona string · lengths	culinary_persona string · lengths	family_persona string · lengths	persona string · lengths
 160~185 33.9%	 126~151 39.8%	 138~159 35.4%	 131~151 35%	 147~169 33.3%	 138~158 31.1%	 70~84 46.6%
전기대 씨는 광주 서구의 하역 현장에서 수십 년간 짐을 싣아 풀리며, 지땀대 원리를 이용해 무거운 자재를 효율적으로 옮기는 베테랑의 면모를 보입니다. 동료들 사이에서 가장 빠르고 정확하게 무게 중심을 잡는다는 평을 듣지만, 가끔은 고집스럽게 본인의 옛 방식을 고수하며 젊은 일꾼들과 투닥거리기도 합니다.	전기대 씨는 주말마다 무등산 자락을 느릿느릿 걸으며 맘을 풀리고, 내려오는 길에 단골 목욕탕에서 친구들과 영겨 앉아 정치 이야기를 나누는 것으로 일주일을 마무리합니다. 무릎 관절이 예전 같지 않아 가끔 찔끔거리면서도 산행만큼은 빼놓지 않는 고집이 있습니다.	전기대 씨는 거실 소파에 깊숙이 파묻혀 텔레비전에서 나오는 옛날 가요 프로그램을 보며 젊은 시절의 추억에 젖어드는 시간을 가장 좋아합니다. 가사가 마음에 들면 흥얼거리며 박자를 맞추다가도, 가끔은 세상 돌아가는 이야기에 화가 나 텔레비전 소리를 높이며 두들기곤 합니다.	전기대 씨는 일주일에 한 번 배달 짜장면과 명수육을 시켜 먹는 날을 손꼽아 기다리며, 2주에 한 번은 아내와 함께 동네 고깃집에서 지글지글 구운 삼겹살에 소주 한 잔을 곁들입니다. 자극적인 음식 맛에 길들여져 가끔은 당수가 걱정되지만, 입에 맞는 음식 앞에서는 건강 걱정을 잠시 잊어버립니다.	전기대 씨는 전·월세 아파트에서 평생의 동반자인 아내와 단촐하게 살아가며, 투박한 전라도 사투리로 서로를 챙기는 소박한 애정을 나눕니다. 무척 속한 성격 탓에 다정한 말 한마디는 서툰지만, 아내가 좋아하는 간식을 숨그머니 식탁에 올려두는 방식으로 마음을 표현합니다.	전기대 씨는 광주 서구에서 평생 하역 일을 하며 살아온 78대 가장으로, 투박한 마디에 삶의 흔적이 배어 있는 성실하고 사교적인 인물입니다.	

데이터 구조

- 총 7가지 페르소나 변수: 직업·스포츠·예술·여행·음식·가족·요약
- 총 6가지 페르소나 속성 변수: 배경, 능력치, 관심사, 가치관
- 총 12개 인구통계 변수: 이름, 나이, 성별, 가족, 지역, 직업 등
- 1개 고유 식별자

- 페르소나 속성 변수

cultural_background string · lengths	skills_and_expertise string · lengths	skills_and_expertise_list string · lengths	hobbies_and_interests string · lengths	hobbies_and_interests_list string · lengths	career_goals_and_ambitions string · lengths
 155~180 22.5%	 150~170 11.4%	 86~103 23.6%	 156~186 16.3%	 80~96 31.6%	 127~147 28.5%
광주 서구에서 평생을 보내며 투박하지만 정겨운 전라도 사투리가 몸에 배어 있고, 시장통 사람들과 머물러 확장지알하게 이야기 나누는 것을 즐깁니다. 초등학교 졸업 후 공바로 현장 일에 뛰어들어 몸으로 부딪치며 살아온 세월이 있어, 격식보다는 사람 사이의 의리와 눈치 빠른 태도를 더 중요하게 생각합니다.	수십 년간 하역 현장에서 다져진 감각으로 짐의 무게 중심을 한눈에 파악해 가장 효율적으로 싣아 옮기는 요령이 탁월합니다. 무거운 자재를 옮길 때 허리에 무리가 가지 않게 지렛대 원리를 이용하는 자신만의 노하우가 있으며, 현장 동료들 사이에서 짐 싣기만큼 가장 빠르고 정확한다는 평가를 받습니다.	['적재물 무게 중심 파악 및 효율적 배치', '현장 자체 결속 및 고정 기술', '하역 작업 동선 최적화', '현장 인력 간의 갈등 중재 및 분위기 주도']	주말이면 무등산 자락을 천천히 걸으며 맘을 빼고, 내려오는 길에 단골 목욕탕에서 뜨거운 물에 몸을 담그며 동네 친구들과 정치 이야기를 나누는 시간을 가장 아낍니다. 일주일에 한 번쯤은 집으로 짜장면과 탕수육을 시켜 먹으며 텔레비전의 옛날 가요 프로그램을 시청하는 것으로 일주일의 피로를 푼다.	['무등산 둘레길 산책', '동네 대중사우나 이용', '전통시장 맛집 탐방', '트롯트 프로그램 시청', '가족과 함께하는 경주나 부여 역사 유적지 여행']	큰 욕심 없이 지금처럼 매일 아침 정해진 시간에 출근해 일하며 건강을 유지하는 것에 만족합니다. 무리하지 않는 선에서 하역 일을 계속하며 아내와 함께 소박하게 생활할 수 있는 생활비를 꾸준히 마련하는 것이 현재의 가장 큰 목표입니다.

3. 한국 synthetic population은 어떻게 만들었나

공식 통계·행정자료를 PGM으로 구조화하고, LLM이 자연어 페르소나를 합성한다.

사용된 데이터

국가 데이터처(성별·지역·산업·직업·여행·여가 인구조사),
대법원(출생연도·성별·이름 데이터), 국민건강보험공단(건강검진정보),
농촌경제연구원(식품소비행태), 네이버 클라우드

1

OCEAN Big Five

개방성(O), 성실성(C), 외향성(E), 친화성(A), 신경증(N) score를 샘플링
고 라벨/설명으로 변환. 각 성격 스코어는 정규분포를 따른다.

2

Demographic Probabilistic Graphical Model

공식 통계와 인구통계 변수들로 만든 PGM 으로 인구통계 데이터 샘플링.

3

Persona attributes

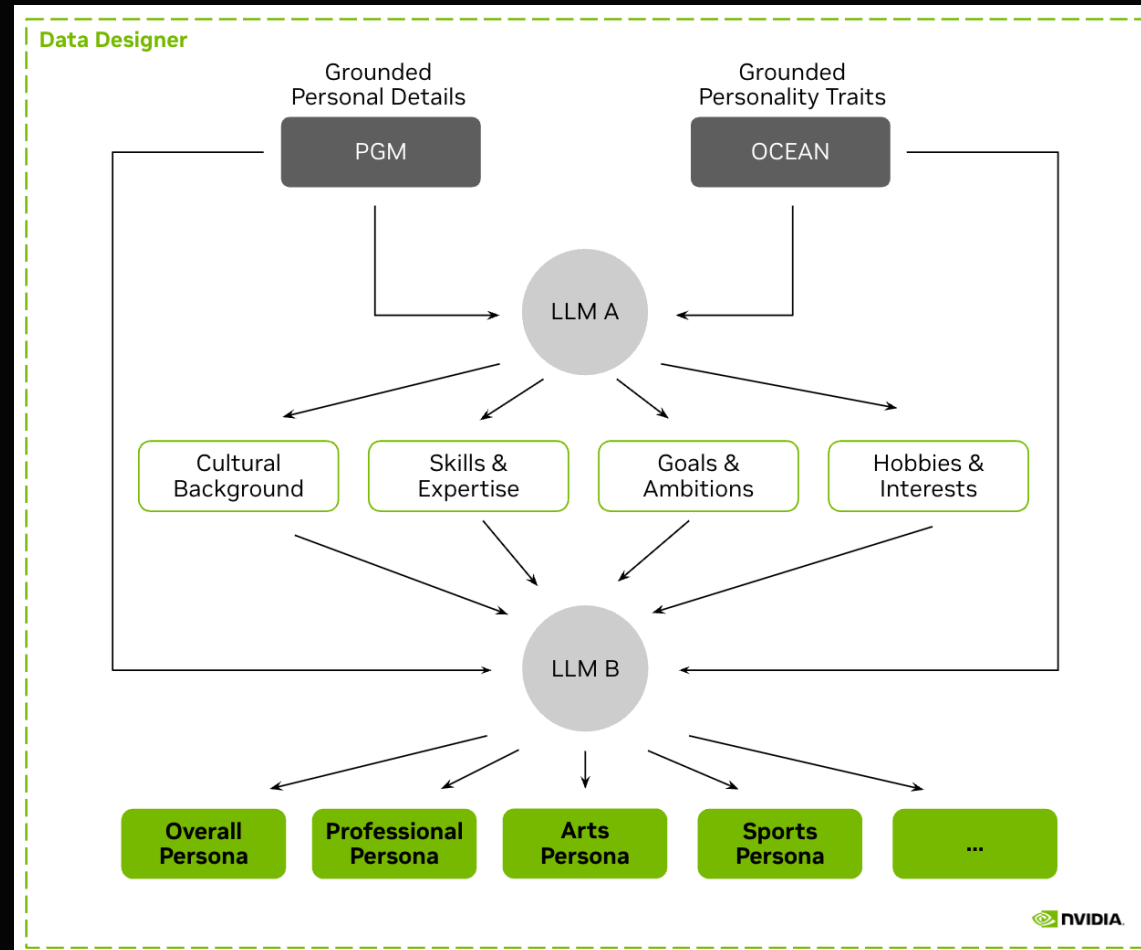
샘플링 된 데이터를 LLM에 입력으로 넣어 사람마다 문화 배경·역량·목표·취미를 설명하는 자연어 데이터 생성.

4

Persona descriptions

PGM/OCEAN 데이터와 페르소나 속성을 LLM에 입력으로 넣어 도메인별
자연어 페르소나를 생성.

PGM은 인구 통계 데이터를 샘플링하고,
LLM은 그 사람의 관점·행동을 자연어로 생성한다.



4. 공개된 PGM 제작 방법

04

공개 프레임워크: NVIDIA-NeMo/SDG-PGMs

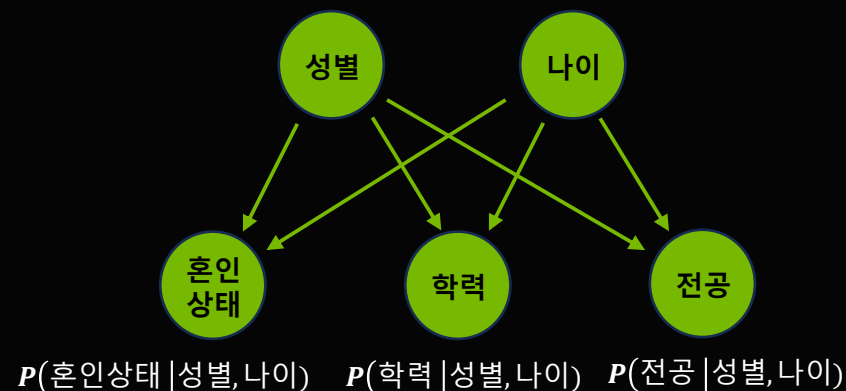
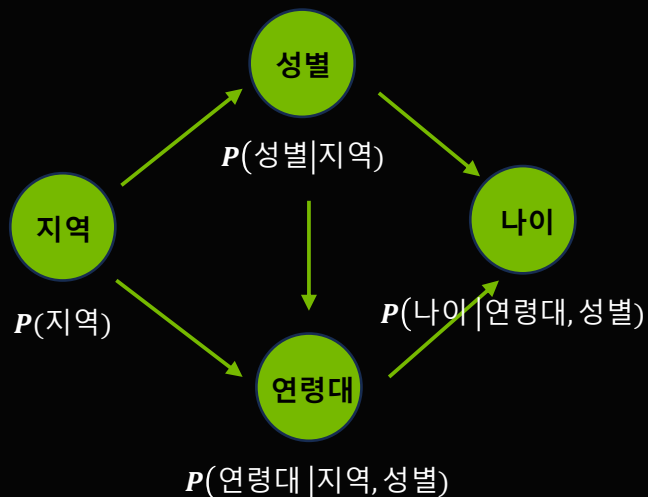
1 데이터 준비
각 분포를 count DataFrame
으로 정리

2 변수·그래프 선언
변수와 Bayesian network 구조
정의

3 Conditional Probability
Distribution 생성

4 샘플링·후처리

5 Data Designer 연결



5 사용 예시

NVIDIA. *Nemotron 3 Super: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning*. 2026.

① 장문맥 모델 학습 데이터 생성

- 장문 학습에는 긴 문맥이 필요하지만, 무관한 문서를 단순히 이어붙이는 것은 적절하지 않다.
- 합성 데이터를 이용해 여러 사람을 비교·검색·집계하는 자연스러운 장문맥 문제를 만들어 학습에 사용할 수 있다.

② 다양한 agent task 생성

- LLM을 통해 합성데이터의 사람이 요구할 task 를 만들어 낸 다음 agent가 이를 수행하도록 학습.
- 다양한 페르소나가 존재해 다양한 task를 만들어 낼 수 있다.
- Agent가 어떤 tool을 사용하는지까지 다양하게 학습할 수 있다.
- NVIDIA는 이 파이프라인을 이용해 약 150만 개의 다양한 tool-calling 방식을 생성했다고 보고.

Persona

62세 은퇴 전기기사
시카고 거주
가격 비교를 중요하게 생각함
스마트폰 사용에는 익숙하지만 복잡한 앱을 선호하지 않음
배우자와 함께 여행

Task

“다음 달 배우자와 보스턴에 3일 다녀오려고 합니다. 오전 출발 항공편과 역에서 가까운 숙소를 찾아 주세요. 총비용은 1,200달러를 넘기지 않았으면 하고, 일정표에도 추가해 주세요.”

Tool-Calling

1. 달력에서 가능한 날짜 확인
2. 항공편 검색
3. 숙소 검색
4. 비용 합산
5. 예산 초과 여부 확인
6. 일정 등록
7. 사용자가 이해하기 쉬운 말로 결과 설명

Thank You!