



# 생성형AI와 양질 데이터 공급의 중요성

2024.3.28.  
금융AI데이터센터



## 목차

- 1 **생성형 AI에서 데이터의 중요성**
- 2 생성형 AI의 활용과 데이터 확보
- 3 생성형 AI와 신용정보의 활용

# 1 Data-Centric AI: 생성형 AI의 궁극적인 성장 방법

“AI에게 데이터란 사람의 음식과 같다” – Andrew Ng



“Data is Food for AI”

앤드류 응(Andrew Ng)

- (전) 구글브레인 설립 멤버
- (현) 스탠포드대 겸임교수



**Action**

- (음식) 음식 조리
- (AI) 모델 훈련



80%

20%



**Preparation**

- (음식) 좋은 재료를 찾아 다듬기
- (AI) 좋은 데이터를 찾아 다듬기

## 2 생성형AI의 Customizing 수준 별 데이터 중요성

생성형 AI를 목적에 맞게 구성·최적화하는 방식은 다양해지고 있으나, 데이터는 언제나 중요함

구분	사전 훈련 Pretraining	파인 튜닝 Fine Tuning	검색 증강 생성 Retrieval Augmented Generation	프롬프트 엔지니어링 Prompt Engineering
개념	모델이 일정한 성능을 출력할 수 있도록 미리 데이터 투입, 훈련	특정 목적에 맞게 모델 일부분에 데이터 추가하여 학습	답변을 만들기 전에 모델 외부의 지식 베이스를 참조	원하는 결과가 나올 수 있도록 질문 가이드 제공
모델의 목적 적합성 확보 대상	<b>모델 전체</b> <ul style="list-style-type: none"> <li>신경망 구조</li> <li>어휘 크기</li> <li>컨텍스트 길이</li> <li><b>훈련 데이터</b></li> </ul>	<b>모델 일부</b> <ul style="list-style-type: none"> <li>특정 작업 튜닝</li> <li><b>주제 별 훈련 데이터</b> 추가</li> </ul>	<b>모델 참조 영역</b> <ul style="list-style-type: none"> <li><b>사용자 데이터</b> 이용</li> <li>검색, 참조 가능한 DB 구성</li> </ul>	<b>모델 프롬프트</b> <ul style="list-style-type: none"> <li>프롬프트 커스터마이징 및 최적화</li> <li>Few-shot Learning (<b>예시 데이터</b> 제공), 페르소나 부여 등</li> </ul>
훈련 시간 및 비용	매우 높음	다소 높음	낮음	낮음
머신러닝 전문성 요구수준	높음	중간	중간	낮음



### 3 생성형AI의 성능과 학습 데이터의 양

AI가 우수한 성능을 출력하기 위해서는, 방대한 양의 학습 데이터 확보가 필수적

#### Data-Centric AI

- ✓ Andrew Ng 교수를 필두로, 세계적 AI 전문가와 기업체 모두 '모델 자체의 구조보다 학습되는 데이터의 분량과 품질이 AI 성능을 높인다'는 의견이 중론
- ✓ 데이터의 품질이 다소 떨어지더라도, 다양성과 분량이 충분하면 어느 정도 보완 가능

초거대 LLM에서 구체적 목적 특화 SLM 활용으로 트렌드 이동

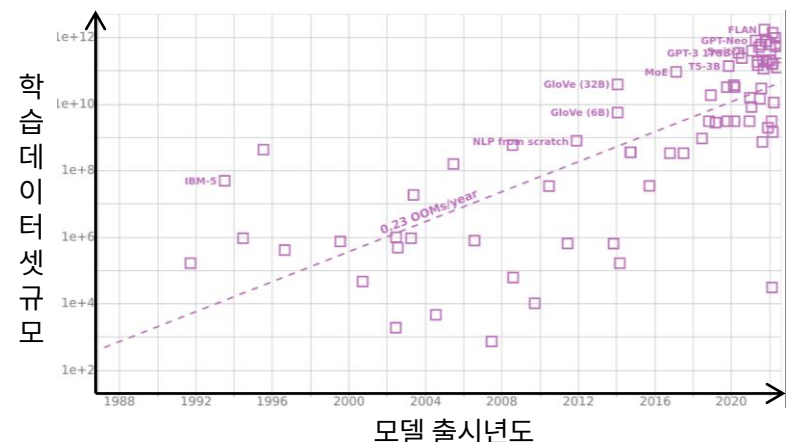
→ 모델 소형화·목적 세분화될수록 다량의 데이터 학습을 통해 목적에 맞는 산출물 확보 필요!

#### 데이터 양의 중요성(Tesla AI 총괄이사)



"필수 준비물: **1백만 차량의 데이터** 엔진, 최고의 AI팀, 그리고 슈퍼컴퓨터"  
(Necessary ingredients include: 1M car fleet data engine, strong AI team and a Supercomputer)

#### 모델 출시년도별 학습데이터셋 규모



## 4 생성형AI와 학습 데이터의 품질

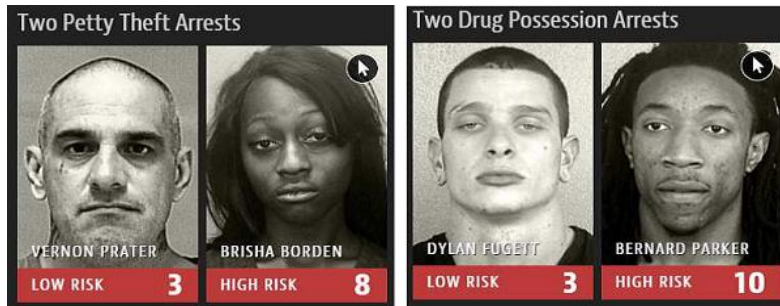
"Garbage in, Garbage out" AI의 불편향성·신뢰성을 위해 데이터 품질 확보 필수

모델 중심 개선 vs. 데이터 중심 개선

	철강 결함 탐지	태양전지 결함 탐지	표면 검사
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

- ☑ 데이터 중심 AI(Data-Centric AI)의 중요성 강조
- ☑ 실무 프로젝트에서는 모델 중심 개선보다 **데이터 중심 개선**이 성능 향상에 훨씬 좋은 효과

AI학습 편향성 예시



- ☑ 뉴스 기사, SNS 문구 등을 학습한 AI에 범죄자 사진으로 재범 위험을 예측하게 하자, 실제 통계와 달리 흑인 재범률을 더 높게 예측

OpenAI GPT-3의 데이터 관련 사전학습

Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Common Crawl (filtered)	<b>410B</b>	-	0.71	1:1.9	<b>570</b>
WebText2	19B	25% > WebText	0.38	1:2.6	50
Books1	12B	Gutenberg	0.57	1:1.75	21
Books2	55B	Bibliotik	0.54	1:1.84	101
Wikipedia	3B	See RoBERTa	0.26	1:3.8	11.4
Total	<b>499B</b>				753.4GB

Table 8. GPT-3 Datasets. Disclosed in **bold**. Determined in *italics*.

- ☑ GPT-3은 데이터를 **재정비**하여 사전학습을 추가 진행함으로써 불편향성·신뢰성을 높임  
→ 양질의 데이터 투입은 필수!

## 5 데이터 품질 향상 요건

우수한 AI를 만들기 위한 데이터 품질의 6가지 요소: **일관성, 정확성, 독창성, 피드백, 포괄성, 밸런스**



- ✓ 앤드류 응 교수는 **좋은 데이터**의 요건을 '일관되게 정의되고, 모델 활용 시 **예상 가능한 모든 경우를 포괄하며, 적절한 피드백을 통해 개선 가능할 것**'으로 규정
- ✓ 그 밖에도 중복되지 않는 **독창성**, 편향되지 않는 **균형성** 등 다양한 품질 요건이 적용

- ✓ 금융 AI 신뢰성 제고를 위해  
특히, **포괄성**과 **균형성(밸런스)**의 중요성이 증대  
→ **편향되지 않은 다양한 소스의 데이터를 활용할 필요**

Data Collaboration 필요성 ↑



## 목차


- 1 생성형 AI에서 데이터의 중요성
- 2 생성형 AI의 활용과 데이터 확보
- 3 생성형 AI와 신용정보의 활용


## 6 금융데이터의 특수성 보완: 금융 말뭉치 공동 구축


금융 말뭉치를 공동 구축하여 한글 기반 전문용어가 많은 금융 데이터에 최적화된 LLM으로 보완 가능

### ④ (금융데이터의 특수성) ① 한글 기반 ② 금융 전문용어

(예시) 일반 도메인으로 학습된 LLM이 금융 전문용어인 '차주'에 관련된 질문에 제대로 답변하지 못하는 모습

 You  
차주의 의미가 뭐야?

 금융권에서 '차주'의 의미가 뭐야?

 ChatGPT  
"차주"란 말은 주로 자동차 관련 분야에서 사용되며, 자동차의 소유주 또는 운전을 주  
담당하는 사람을 의미합니다. 일반적으로 자동차를 소유하고 있는 개인이나 법인을 가  
리키며, 자동차를 구매하거나 임대한 사람을 지칭할 때 사용됩니다. 예를 들어, 자동차  
대출이나 리스 계약을 체결할 때 차량의 '차주'가 계약의 당사자로 등장하게 됩니다.

Llama 2 70B



"차주"는 금융권에서 사용되는 용어입니다. 이 용어는 "次の世代" (Next Generation)의 약자로, 일반적으로 기업의 경영 또는 리더십을 이어받을 차  
리더들을 指します. 차주는 기업의 장기적 성공을 위해 필요한 리더십 및 기  
술 능력을 갖추고 있는 사람들입니다. 차주 개념은 기업의 지속적 성공을 위  
해 필요한 리더십 및 기술 능력을 갖추고 있는 사람들을 육성하는 것을 목표  
로 합니다.

### 금융 말뭉치(Corpus)

- 금융 전문용어데이터 집종을 통해 **금융 도메인에 특화된 LLM 생성 및 금융권 공동 활용 가능**
- 정보보호, 영업기밀 등 이슈로 금융사별 내부 데이터 집중에 어려움 존재

→ 금융회사 외부 데이터 추가 수집·활용 검토

- ① 금융협회 보유 데이터, ② 금융 관련 법령 정보, ③ 금융 관련 교육 데이터  
(가이드라인, 모범규준 등) (금융연수 교재 등)

## 7 LLM 환각현상 보완 방안: RAG DB 공동 활용

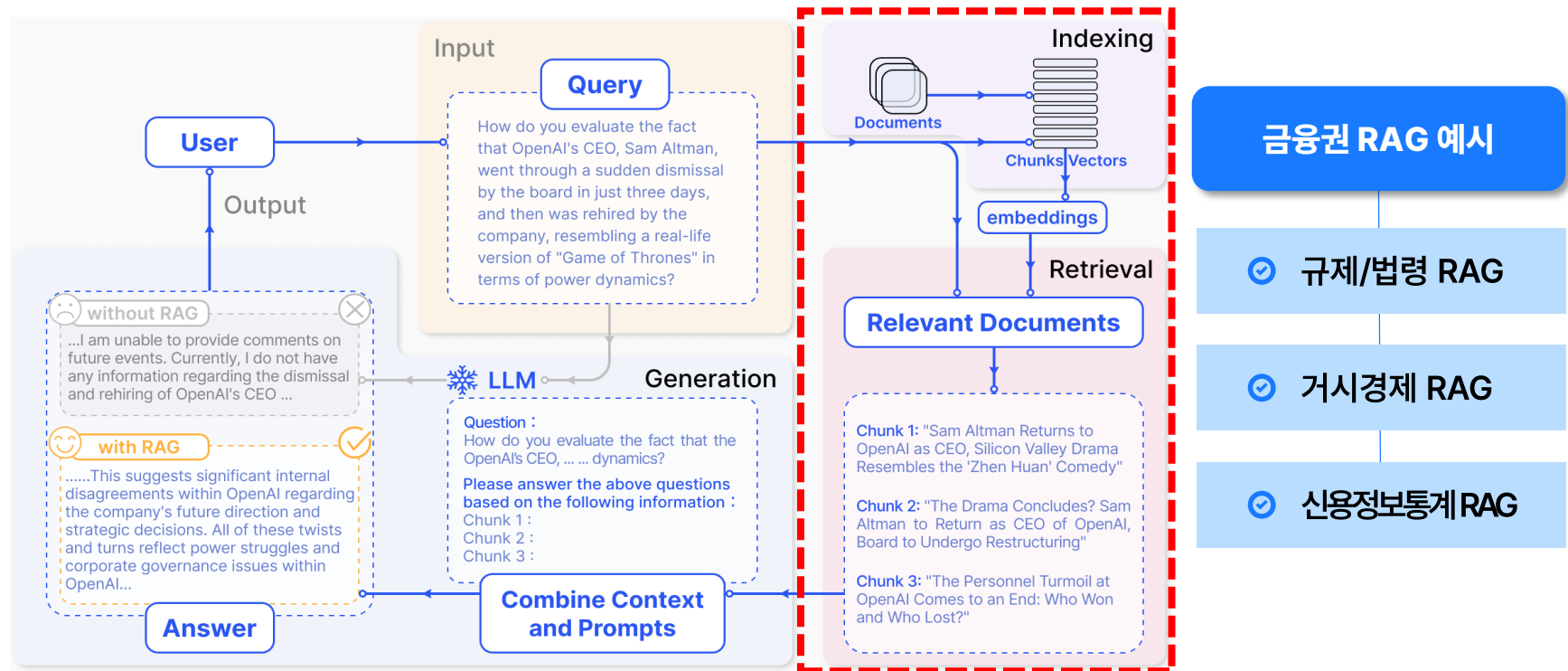
금융권을 위한 환각현상 보완기술(RAG DB)을 구축 및 공동 활용하여  
효율적으로 문제 보완 가능

### 검색증강생성(RAG)

Retrieval Augmented Generation

LLM 출력을 최적화하여 응답을 생성하기 전에 외부 소스에서 가져온 최신 정보를  
참조하는 프로세스로 환각현상\*을 보완하여 생성 AI 모델의 정확성과 신뢰성을 향상시킴

\* (환각현상) 사실이 아닌 정보를 그럴듯하게 대답하는 현상으로 전문지식 분야일수록 발생확률 증가



### 금융권 RAG 예시

☑ 규제/법령 RAG

☑ 거시경제 RAG

☑ 신용정보통계 RAG

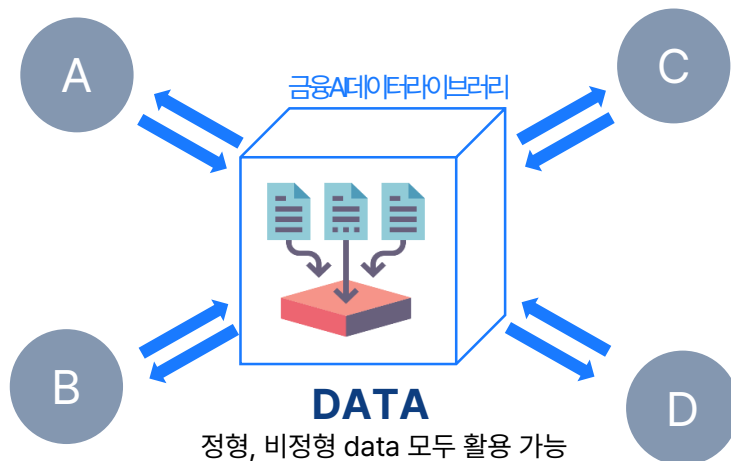
## 8 데이터 확보 방안: 데이터·모델 공동 활용

데이터 공동 활용 또는 연합학습을 통해 각 금융회사의 데이터 부족 문제를 보완

### Data 공동활용(표준데이터셋)

표준데이터셋을 구축하여 다양한 기관간 자사 보유  
**데이터**를 결합·저장·공동 활용하여 데이터 활용 범위를  
확대 가능(데이터 표준화를 통해 품질 제고)

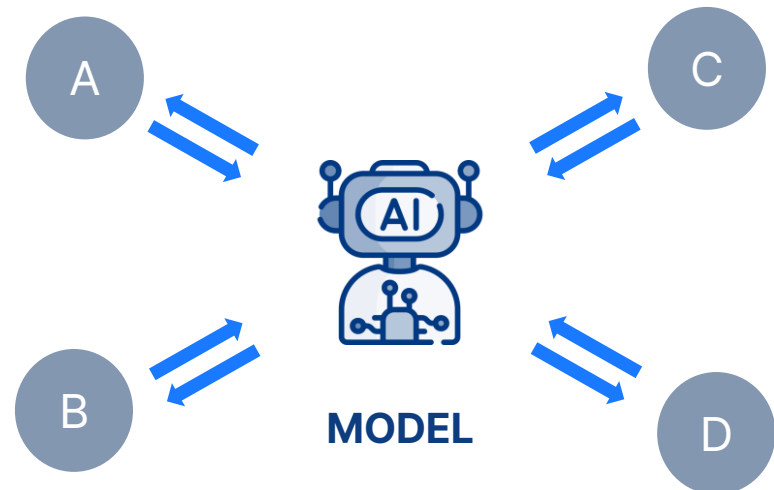
(예시) 금융AI 데이터 라이브러리 활용 업권 별 데이터 결집,  
업권 공동 FDS 모델 개발



### Model 공동활용(연합학습)

각 기관 내에서 데이터 학습 후 **모델**을 공유하는  
연합학습(Federated Learning) 활용 가능

(예시) Gaia-X: 독일 개별 자동차들의 주행 정보를 모아  
Central AI model을 업데이트하여 활용





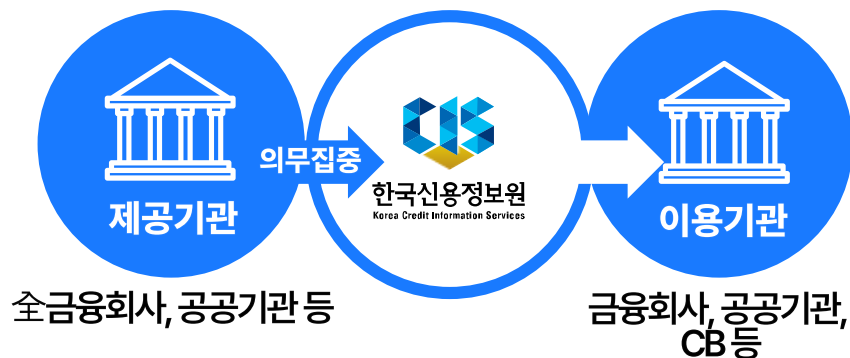
## 목차

- 1 생성형 AI에서 데이터의 중요성
- 2 생성형 AI의 활용과 데이터 확보
- 3 생성형 AI와 **신용정보**의 활용



## 9 생성형AI 활용을 위한 신용정보 품질 관리

신용정보(일반·보험·기술)는 생성형 AI에 활용 가능한 **포괄적**(全국민), **양질**의 데이터로서  
종합신용정보집중기관인 신용정보원으로 집중되어 안전하게 통합 및 품질관리 中



일반신용정보

대출, 연체, 신용카드 개설, 세금체납 등

보험신용정보

보험계약, 청구, 사고, 지급 등

기술신용정보

기술신용 대출, 기술평가 등

집중 일반신용정보(예시)

개인대출정보 DB	
기준년월	
차주일련번호	
차주구분코드	
업권코드	
기관일련번호	
대출상품코드1	
대출상품코드2	
대출상품코드3	
대출년월	
대출금액	
금리등록여부	
상환방식	
만기월	
대출금리	
연간원리금상환금액	

업권코드

01	국내은행
05	신용카드사
05	상호저축은행

대출상품코드1

0031	개인대출
0037	장기카드대출

대출상품코드2

150	신용>학자금
000	카드대출
240	담보>보증서담보
220	담보>주택담보

대출상품코드3

200	햇살론
000	(서민금융아님)

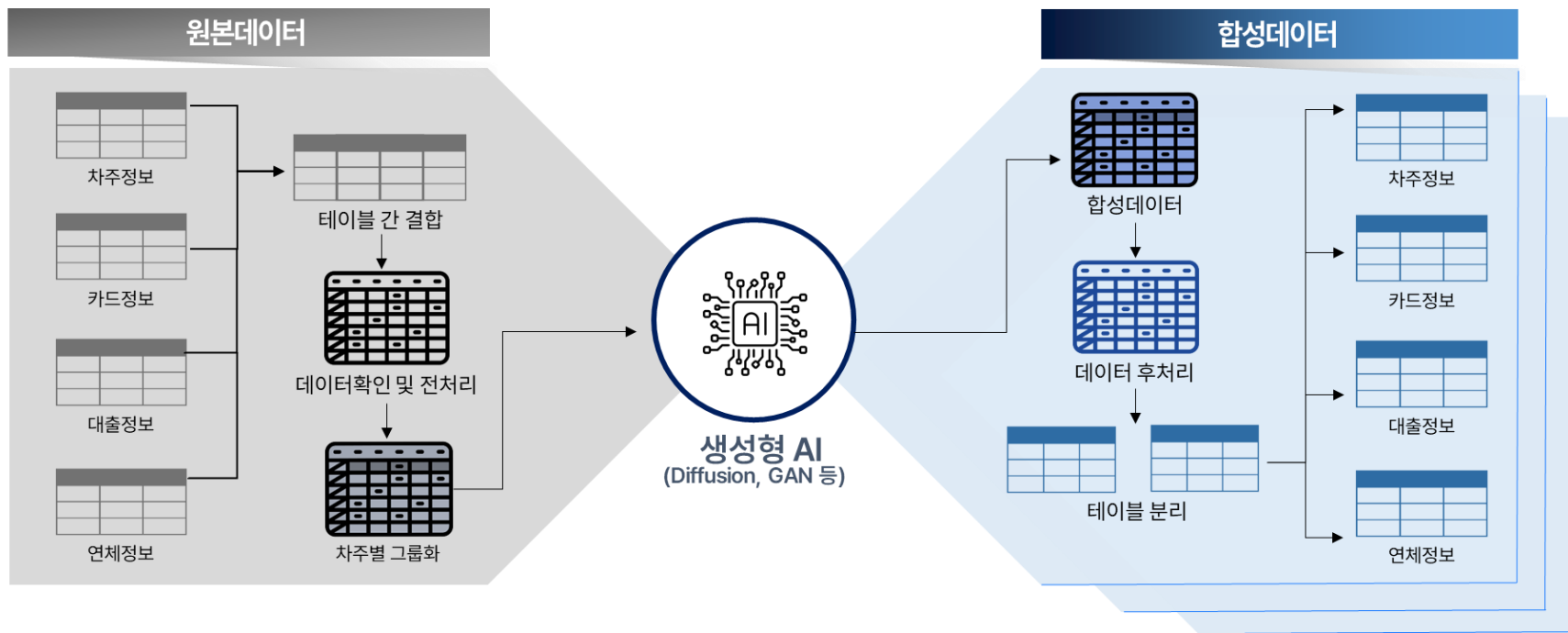
상환방식

02	원금균등분할상환
03	거치식-원금균등분할상환
04	원리금균등분할상환

## 10 신용정보를 활용한 합성데이터

금융기관 및 데이터 자원이 부족한 스타트업 등이 AI 모형 학습 등에 활용할 수 있도록  
신용정보 합성데이터를 개발하여 제공

🕒 (합성데이터) 실제 데이터의 통계적 속성을 갖지만 실제 정보를 포함하지 않는 생성형 AI 기술 기반의 데이터

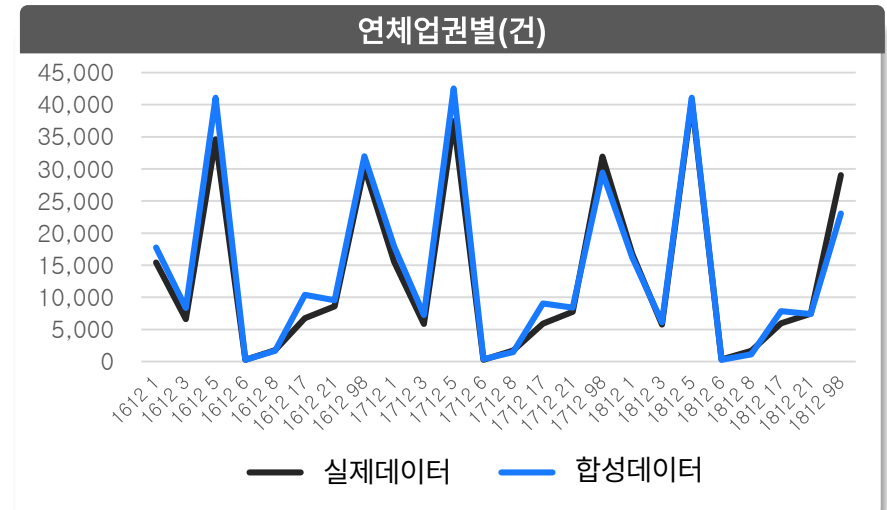
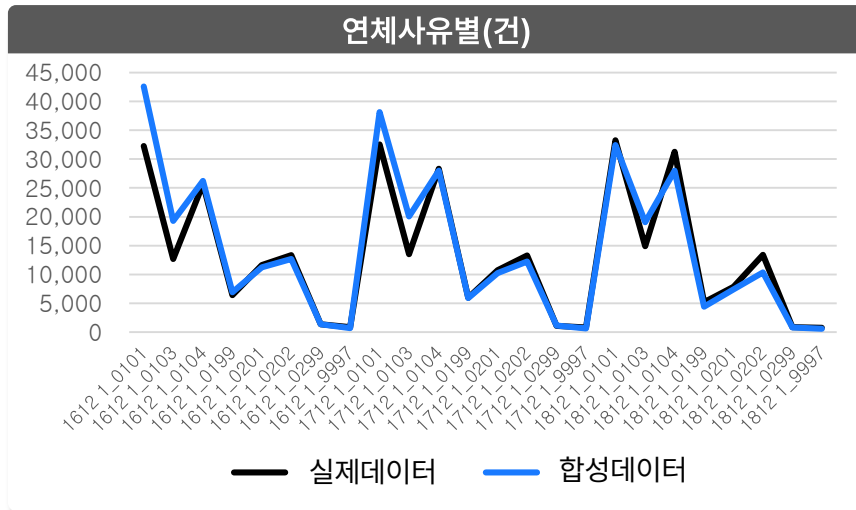


합성데이터로 **증강**된 데이터를 통해 데이터 불균형(Imbalanced Data) 이슈 해소 가능

## 참고 신용정보를 활용한 합성데이터(예시)

금융기관 및 데이터 자원이 부족한 스타트업 등이 AI 모형 학습 등에 활용할 수 있도록  
신용정보 합성데이터를 개발하여 제공

🕒 (통계적 유사성) 연체사유 및 연체업권 별 실제-합성 연체건 데이터를 비교한 결과, 유사한 패턴을 보임



🕒 (목적 별 합성데이터 개발) 모형개발 또는 시계열 분석 등 특정 목적 별 신용정보를 활용한 합성데이터 개발 가능

모형 개발용  
합성데이터

특정 모형개발(개인신용평가, 기업신용평가 등)을 목적으로 신정원의 신용정보 데이터를  
차주별로 조합하고 요약항목을 생성 후 이를 바탕으로 합성데이터 개발

시계열 분석용  
합성데이터

연체여부 및 대출잔액, 보유보험수 등을 시계열로 합성하여 제공

## 11 신용정보원의 AI·데이터 서비스

신용정보원은 전 데이터 수명주기에 걸친 AI·데이터 서비스를 제공하여  
금융 데이터 Hub로서 금융권의 AI·데이터 활용 제고 역할에 힘쓰겠습니다.





**감사합니다**

