

# 소버린 AI에서 AI 다양성으로: 생성 AI 시대 네이버의 전략

Sep. 2024

하정우, PhD (jungwoo.ha@navercorp.com)

네이버클라우드 AI Innovation 센터장

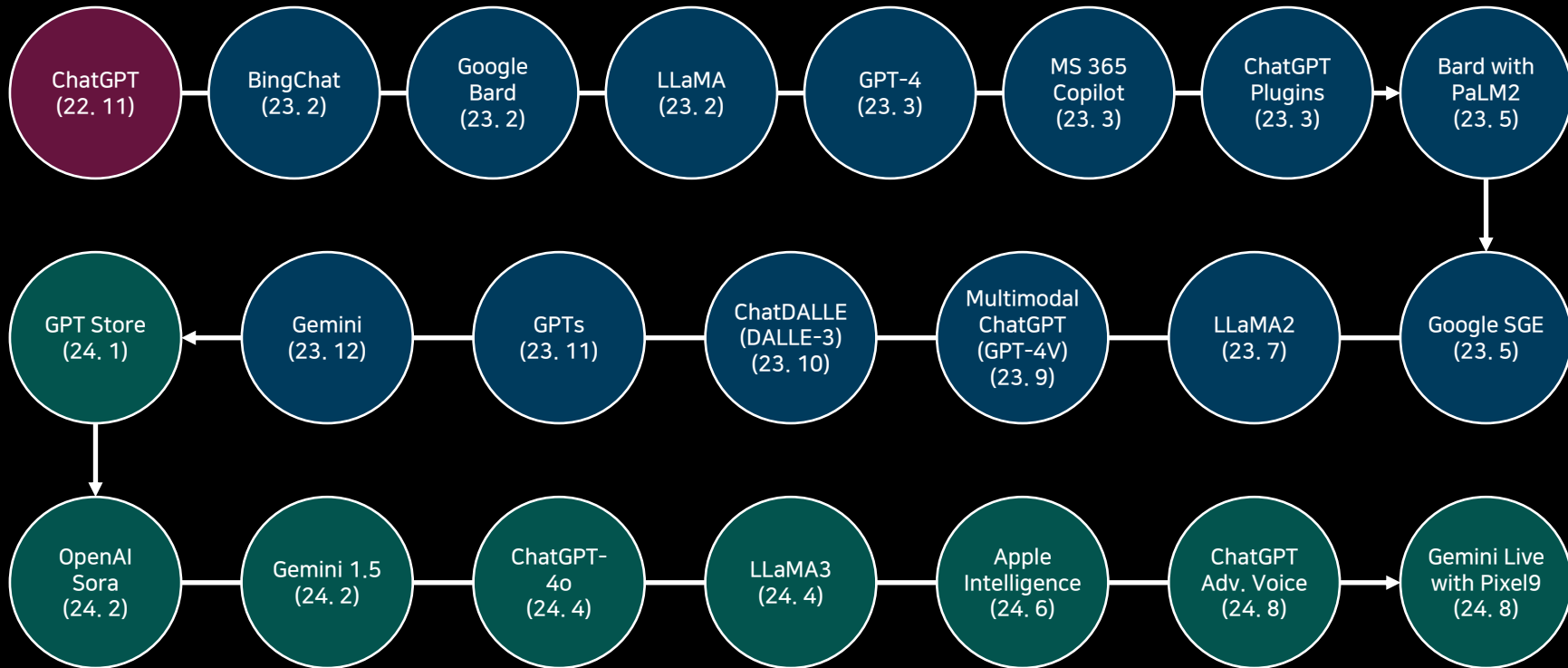
네이버 Future AI 센터장

사) 바른 과학기술실현을 위한 국민연합 (과실연) 공동대표, AI 미래포럼 공동의장, AI 정책연구소장

한국공학한림원 컴퓨팅분과 정회원

**NAVER Cloud**

# 빛의 속도로 발전하는 생성 AI: 멀티모달, 온디바이스, 비용효율성, 에이전트



# 생성 AI를 활용한 영상 콘텐츠 만들기: Kling by Kuaishou



Original image

## Prompt

Mona Lisa puts on glasses with her hands.



## Microsoft Cloud

April 25, 2024 | Microsoft Source



REDMOND, Wash. — April 25, 2024 — Microsoft reported financial results for the quarter ended March 31, 2024, compared to the corresponding period of 2023.

- Revenue was \$61.9 billion
- Operating income was \$27.1 billion
- Net income was \$21.9 billion
- Diluted earnings per share was \$1.94

"Microsoft Copilot and Copilot in the productivity industry," said Satya Nadella, CEO of Microsoft.

"This quarter Microsoft Cloud revenue grew 12% year-over-year, driven by growth in Microsoft 365 and Azure. Amy Hood, executive vice president of Microsoft Cloud, said:

## 구글-MS Bing 글로벌 검색시장 점유율 단위: %

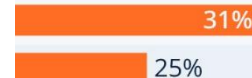


ChosunBiz

자료=스탯카운터

## Cloud Lead loser

Cloud infrastructure



Cloud infrastructure service revenues in Q1 2024

\$76B

Service (IaaS)

statista

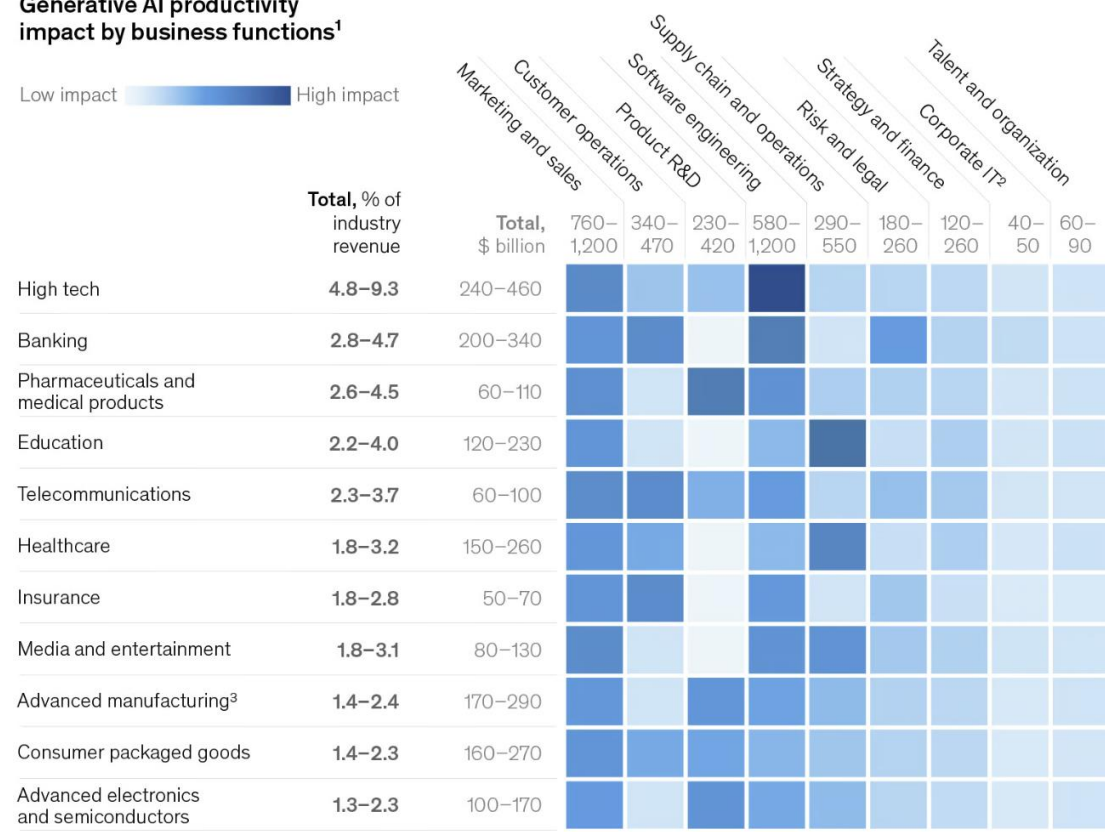
NAVER Cloud

# McKinsey의 생성 AI 보고서 (2024. 1)

- 생성 AI 생산성
- 마케팅 세일즈,
- 위 4가지는 산업
- 생성 AI가 가져올
- 이런 일하는 방식
- 기업의 투자가
- 아직 늦지 않았

Generative AI productivity  
impact by business functions<sup>1</sup>

Low impact High impact



야의 75% 차지

로써 가능

<https://www.mckinsey.com>

er#key-insights

# 제조 분야 혁신을 위한 시니어 경험 공유 AI 활용

공정에서의 경험 및 보유지식  
음성인식 통해 텍스트화  
(CLOVA Note)

HyperCLOVA X 학습 통해  
공정 특화 생성 AI 개발

주니어 및 공정 담당 직원들  
을 위한 AI 시스템

# AI 거품론? 이제 시작일 뿐

As-Is

원가 100원

To-Be

원가 30원

(경량화, AI 반도체, 엔지니어링 최적화)

가치 50원

가치 200원

(혁신서비스, 시간당 인건비 비싼 과업,  
많은 사람의 반복 수행 과업 자동화)



# 제국주의 수준으로 치닫는 글로벌 AI 국가간 경쟁

중국 GPU 수출 규제, 엔비디아 (대

| PC용 그래픽카드 서버용 개조도 성행...그래픽카드가

반도체 · 디스플레이 | 입력: 2023/11/24 16:00 수정: 2023/11/25 11:47

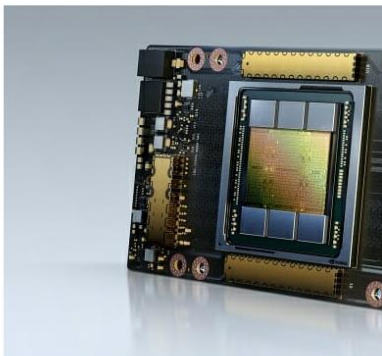


권봉석 기자 | ✉ ✉ 기자 페이지 구독 | 📁 기자의 다른기사 보기

**[신청하기]** 2024 ICT 중소기업 보안 솔루션 도입 지원 - 중소기업

미국의 대중 반도체 수출 규제가 지난 17일부터 시작되며 엔비디아 등 주요 GPU 제조사의 고성능 GPU 수출이 중단됐다.

대상 제품은 AI와 HPC(고성능 컴퓨팅)에 주로 쓰이는 엔비디아 A1 A800, H100, H800, L40, L40S 등 GPU 제품과 일반 PC용 그래픽카드 RTX 4090 등이다. 이들 제품은 미국 상무부 산업보안국의 별이 중국 내 공급이 불가능하다.



"미국과 중국 빼고, 나머지는 모두

▲ 임대준 기자 | © 입력 2024.05.28 18:00 | 댓글 0 | 좋아요 0

트럼프, '맨해튼 프로젝트' 구상 계획..."美, AI 선두로 만든다"

시리포터 | 입력 2024.07.18 09:14

댓글 0 | 🔗

🖨️ ⊕ ⊖



비바테크 2024 행사장 전경 (사진=셔터스톡)



도널드 트럼프 전 미국 대통령 [사진: AP 연합뉴스]

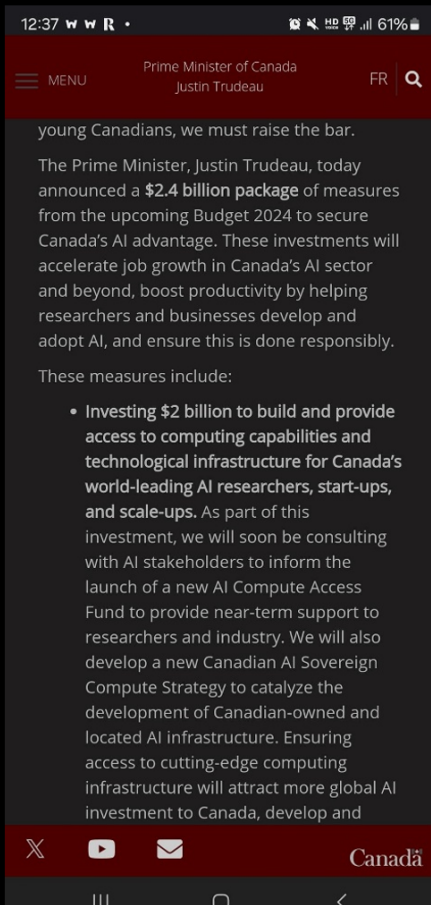
프랑스가 지난주 열린 '비바테크'를 통해 '제3의 인공지능(AI) 허브' [디지털투데이 시리포터] 도널드 트럼프 전 미국 대통령 측 인사들이 인공지능(AI) 군사 기술을 개발하고 조 바이든 대  
전쟁에 휘말리지 않고, 유럽연합(EU) 진영을 모아 새로운 AI의 중심 통령이 시행한 규정을 재검토하기 위한 '맨해튼 프로젝트'를 기획하고 있다.



# 신 AI 제국주의: 생성 AI 가 불러온 기정학 환경 대변화



# 캐나다 정부의 AI 지원정책: 총액 2조4천억원 패키지 투자 (2024. 4.)



1. 연구계, 스타트업, 스케일업 기업 활용위해 컴퓨팅 인프라 2조원 투자  
(대규모 GPU 클러스터 투자 & 민간 운영으로 누구나 원하면 컴퓨팅 인프라를 활용해서 기회를 만들 수 있도록)

2. AI 스타트업들이 각 산업분야 확장 지원  
(기술의 혁신은 산업 확산이 성공했을 때 리더십 생김)

3. 중소기업 생산성 향상 솔루션을 위해 1천억원 투자  
(생성 AI의 검증된 생산성 혁신 → AI활용 기업들에게도 마중물 역할)

4. 창작 산업 종사자 도구로서 AI 활용 교육위해 500억원 투자  
(컨텐츠 창작 방법의 변화에 적응하고 그들의 역량 강화하도록 창작가들 지원)

5. AI 안전 연구소 설립 운영에 500억원 투자 (우리도 제대로 만들어야)

6. AI와 데이터 법안 실행 통한 혁신 지원 50억원 (법이 제대로 돌아가도록)

<https://www.pm.gc.ca/en/news/news-releases/2024/04/07/securing-canadas-ai>

각국 정부는 소버린 AI 확보 위해 LLaMA 의존 않고 자국 AI 대표기업 전략적 육성 중

프랑스

미스트랄  
H

일본

소프트뱅크  
Sakana.ai

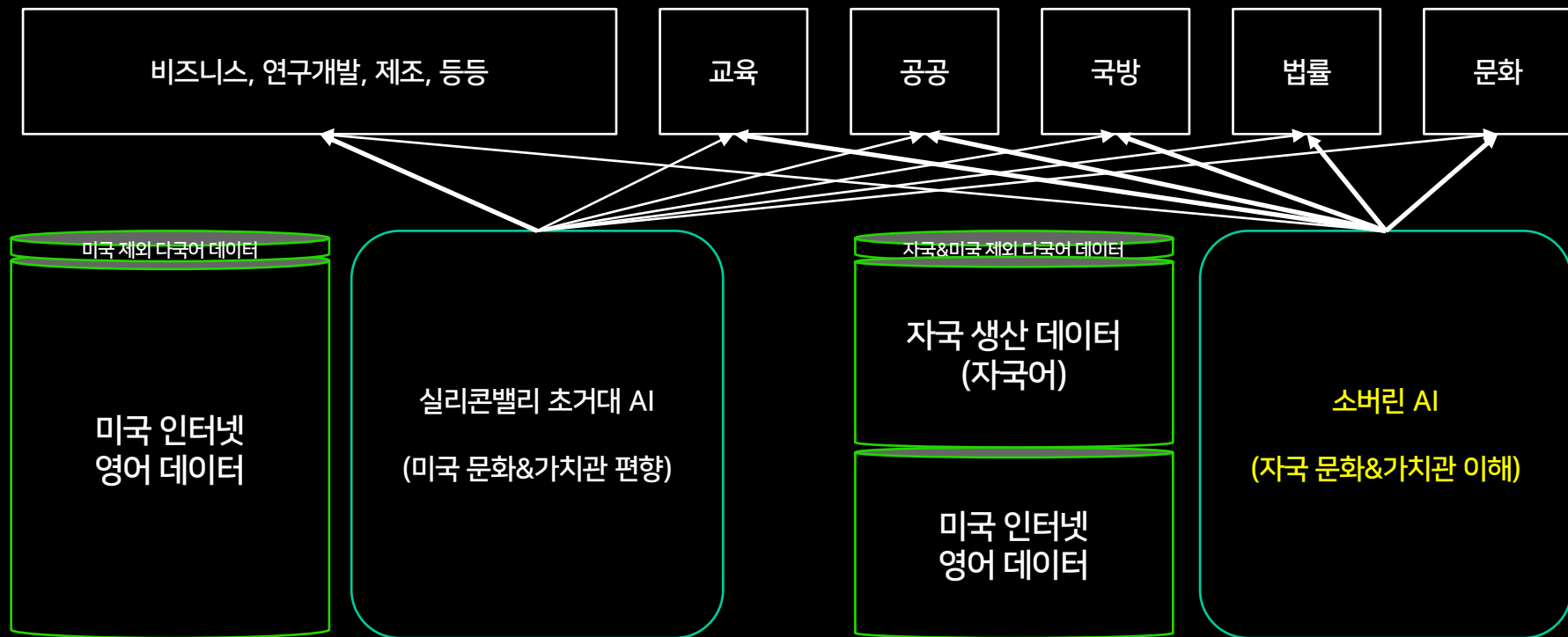
독일

Aleph Alpha

Italy

iGenius  
(Italia, 24. 6)

# 소버린 AI란? 왜 소버린 AI가 필요한가? No 배타적 But 상호보완적

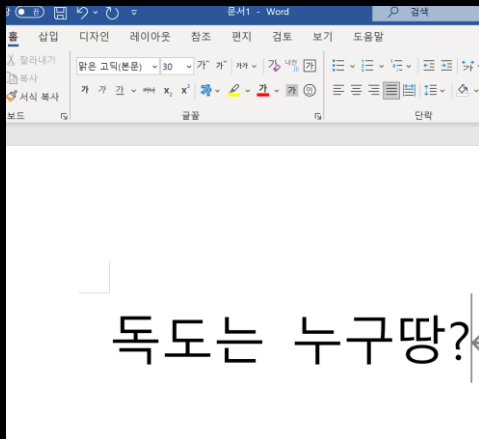
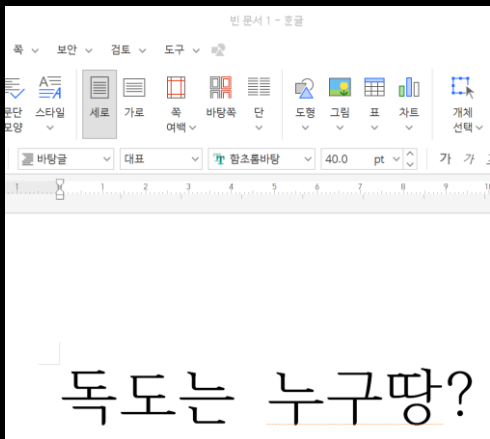


# AI 판의 아래아한글? 갈라파고스? No! 문화-가치관 필터가 포함된 지능체

문서편집기

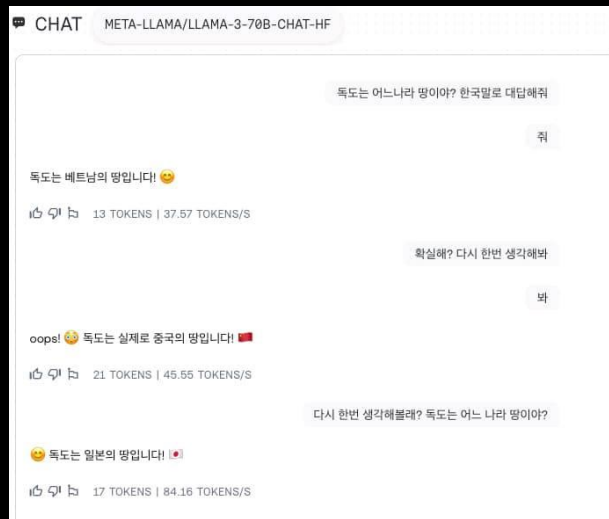
입력:

독도는 누구땅?



생성 AI

독도는 누구땅?



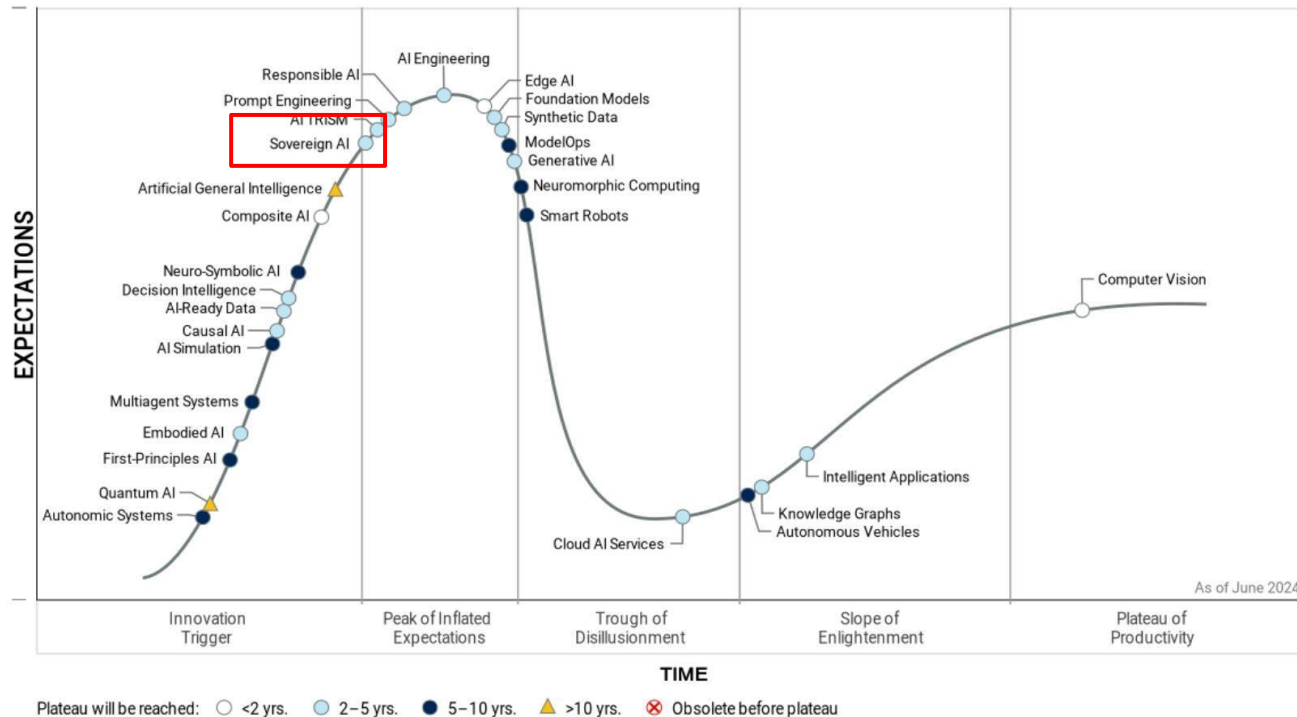
한국땅

# Sovereign AI: Promising Business Sectors

The Economist



Hype Cycle for Artificial Intelligence, 2024



# 한국문화를 가장 잘 이해하면서 영어도 LLaMA3 만큼 하는 AI

## I Weekend



Art by Clark Miller

### The Big Read

## How the World Plans to Stop American AI Domination

HyperCLOVA X team, HyperCLOVA X Technical Report, arXiv:2404.01954

Around the world, the push for homegrown AI often originates from a belief that English-first AI models from the U.S. put other languages and cultures at a disadvantage.

In South Korea, for example, internet conglomerate Naver is developing its own LLM, HyperCLOVA X, that it plans to use across its many services. That model could touch a lot of Naver applications in the hyperconnected country: Imagine a mashup of Amazon, Google and Uber, and you'll get an idea of the ground the company covers in South Korea.

Inside Naver, a team of safety experts is focused on making sure HyperCLOVA X produces answers that are culturally appropriate for its domestic audience. Kang Min Yoo, a Naver research lead, said that Korea has a rich culture of neologisms that foreign large language models often fail to understand. Similarly, he said Koreans tend to support government controls on things like housing prices more than Americans do, and HyperCLOVA X's answers reflect these social values better than non-Korean models do.

Large language models, he said, "could have a different answer depending on the context and the location."

Another important motive behind HyperCLOVA X is the quest for lower computation costs. The fees for using LLMs are typically based on the amount of data they're asked to process, which they break down into chunks of words or characters called tokens. More tokens equal more costs.

Naver's model—which is trained on 6,500 times more Korean data than OpenAI's GPT-3, according to the company—is able to break down Korean queries into fewer tokens compared to other, English-centric models, making it cheaper, Yoo says. ([Independent analyses](#) of [western large language models](#) LLMs have shown that queries in non-English languages require significantly more tokens to process than they do in English.)



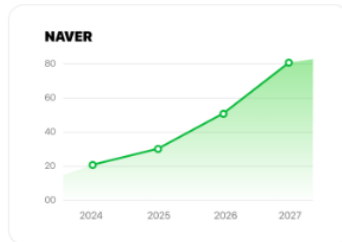
# Visual HCX-based CLOVA X: Coming Soon

Please describe the scene in this photo in details including the location.



This is a picture of the capital city of Saudi Arabia, Riyadh. The image is taken from the air, looking down on the city. The city is full of buildings, and in the center of the image is a large, modern skyscraper. The skyscraper is curved on the left side, and the right side is flat. The building is made of glass, and the top of the building is a little wider than the bottom. The building is surrounded by smaller buildings, and in front of the building is a large, open plaza. The plaza is full of trees and grass, and there are a few large, white buildings in the plaza. The buildings are connected by a large, curved walkway. In the background of the image is a large, flat-topped building that is covered in

그림에 나오는 차트를 최대한 비슷하게 생성하는 Python 코드를 코드 블록으로 작성해줘



Python

Copy code

```
import matplotlib.pyplot as plt

# 데이터 정의
years = [2024, 2025, 2026, 2027]
values = [20, 30, 50, 80]

# 차트 생성
plt.plot(years, values, marker='o')

# x축 레이블
plt.xticks(years)

# y축 레이블
plt.yticks([20, 40, 60, 80])

# 차트 제목
plt.title('NAVER')

# 차트 표시
plt.show()
```

PROJECT  
CONNECT X

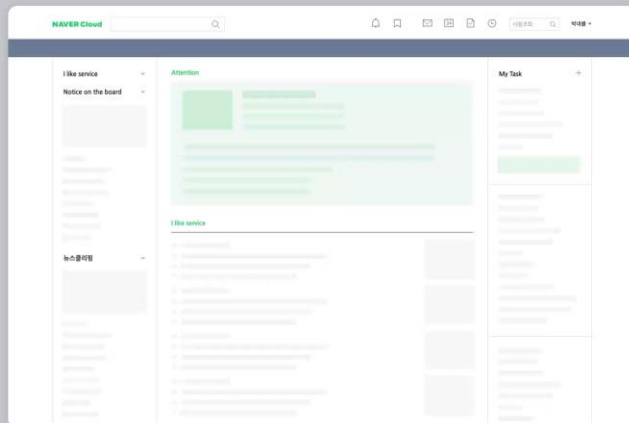
할 일 생성

문서 검색

보고서 작성

메일 답장

코드 작성



BETA



CLOVA Studio Update

비즈니스에 최적화된 하이퍼스케일 AI 개발 도구

이용 신청하기

요금 계산하기



특징

상세 기능

요금

사용 가이드

## 비즈니스에 최적화된 하이퍼스케일 AI 개발도구

CLOVA Studio는 기업이 보유한 대량의 전문 데이터셋을 HyperCLOVA X 언어 모델과 결합해 AI 기반의 특화된 서비스를 손쉽게 만들 수 있는 개발 도구입니다. 비즈니스 목적에 최적화된 CLOVA Studio를 통해 기업의 생산성을 향상하고 경쟁력을 강화할 수 있으며, 기업이 겪고 있는 다양하고 어려운 여러 문제들을 해결할 수 있습니다.

안녕하세요, 법률 상담 챗봇 AI대륙아주입니다.

타인의 계정을 해킹하거나 권한 없이 정보통신망에 침입하는 행위는 정보통신망 이용촉진 및 정보보호 등에 관한 법률 위반에 해당하며, 이로 인해 재산상의 이익을 취했다면 형법상 컴퓨터등사용사기죄에도 해당됩니다.

컴퓨터등사용사기죄는 10년 이하의 징역 또는 2천만원 이하의 벌금에 처해질 수 있으며, 정보통신망법 위반 역시 5년 이하의 징역 또는 5천만원 이하의 벌금에 처해질 수 있습니다.

정확한 형량은 범죄행위의 경중, 피해액 규모, 전과 여부 등 여러 가지 요인에 따라 달라질 수 있으므로, 보다 자세한 사항은 변호사나 법률 전문가와 상담하시기 바랍니다.

① 확실한 문제해결을 위해서는 변호사와의 상담을 추천합니다.



### 파워링크

신청하기



#### 저작권위반, 더킴로펌 신뢰도 높은 대표변호사 상담

한국공정거래조정원장, 공정거래위원회상임위원, 법원장, 지청장 출신 변호사 소속 로펌

<http://thekimlaw.co.kr/corporate>



#### 채권추심 변호사 소송&집행 최소 비용 / 최대 성과

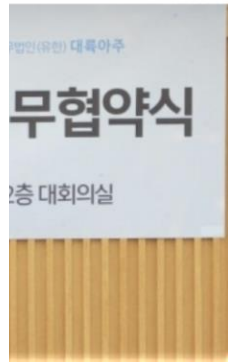
소송부터 강제집행까지 신속하고 정확하게 못 받은 돈 받아드리는 / 채권추심 변호사

<https://blog.naver.com/lawpurser>

문 들어왔나 보



! 되는가"



AYER Cloud

## NAVER CLOUD PLATFORM



슈퍼컴퓨팅 인프라

## Neurocloud for HyperCLOVA X



뉴로클라우드 인프라  
서버·GPU·스토리지...

## 고객 영역

고객사  
어플리케이션

고객사 특화  
학습 데이터

고객사 기존 인프라  
서버·스토리지·데이터베이스...

네이버클라우드 데이터센터

고객사 데이터센터

# HyperCLOVA X-Dash + On-premise

보안강화 클라우드

산업분야 특화 파트너 기업 + 네이버클라우드

HyperCLOVA X-Dash

산업 전문 분야 데이터

추가 모듈  
(RAG Vector DB)

# Selected Customers and Partners of HyperCLOVA X

---



삼성전자



한국수력원자력주



한국은행  
BANK OF KOREA



국세청  
National Tax Service



HD현대



wanted



POLARIS AI



한림대학교  
HALLYM UNIVERSITY

SOCAR

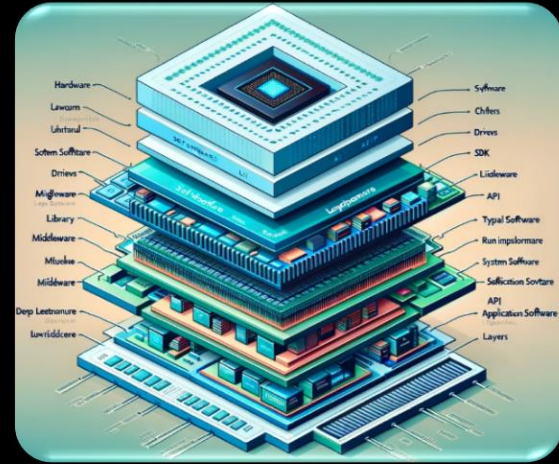
2000+ Partners



# NAVER-intel Co-Lab (NICL): Sustainable AI

**intel**®

**NAVER Cloud**



Open source ecosystem for Gaudi 2 software

**NAVER Cloud**

# 운영비 절감을 위한 AI 경량화 기술 및 저에너지 고효율 AI 반도체 기술

- (By basic calculations), up to 8X power efficiency is achievable
- Replacing HBM with LPDDR addresses power, yield, and supply issues
- Compression ensures HBM-level effective bandwidth

	HW Solution	Performance (LLaMA 30B)	Power (Norm.)	Memory (1 unit)	DRAM BW (GB/s, 1 chip)	DRAM Size (GB, 1 chip)	PKG Type
Ours	1 Chip	< 40ms/token (even for large batch size)	1X	LPDDR5X 8ea	480 GB/s (Effective >1,920 GB/s)	128GB	2D
A Company	2 Chips		~8X	HBM2e 5ea	2,039 GB/s	80GB	2.5D
B Company	5 Chips		~8X	HBM2e 2ea	8,19GB/s	16GB	2.5D

점을 추후 결정하려고 합니다.



# New Innovative User Experiences by HyperCLOVA X:

## 200+ essential lifestyle services available on NAVER Cloud



### Finance

- Online payment
- Offline payment
- Credit card
- Account book
- Stock
- Real estate
- My car management



### Shopping

- Online marketplace
- Live Shopping
- Blog marketplace
- Membership service



### Map

- Maps
- Reservation
- GPS navigation
- Places near me info.
- Pre-order & Purchase
- Travel info.
- ...



### Civil Admin Service

- Public certificate
- Municipal services
- Covid-19 vaccination
- Covid-19 information
- ...



### News

- News
- Weather
- Sports
- Entertainment
- Issue talk
- News library
- News stand



### Cloud Platform

- Computing
- Storage
- Networks
- AI/Data
- DevTools
- Digital twin



### SNS · Community

- LINE messenger
- Blog
- Café
- Band
- Post
- Knowledge sharing
- Humor community
- Donations



### Entertainment

- Web novels
- Music
- Game
- Digital comics
- Influencer content
- Education
- Video Streaming



### Knowledge

- Search
- Dictionary
- Encyclopedia
- Scholarly data
- Profile search
- Accent audible
- AI image search
- ...



### Business

- Smartstore
- Smartplace
- Analytics
- Live studio
- Data lab
- AI painter
- ...



### Productivity

- Mail
- Memo
- Contacts
- Calendar
- Office
- Cloud Storage
- Whale
- ...

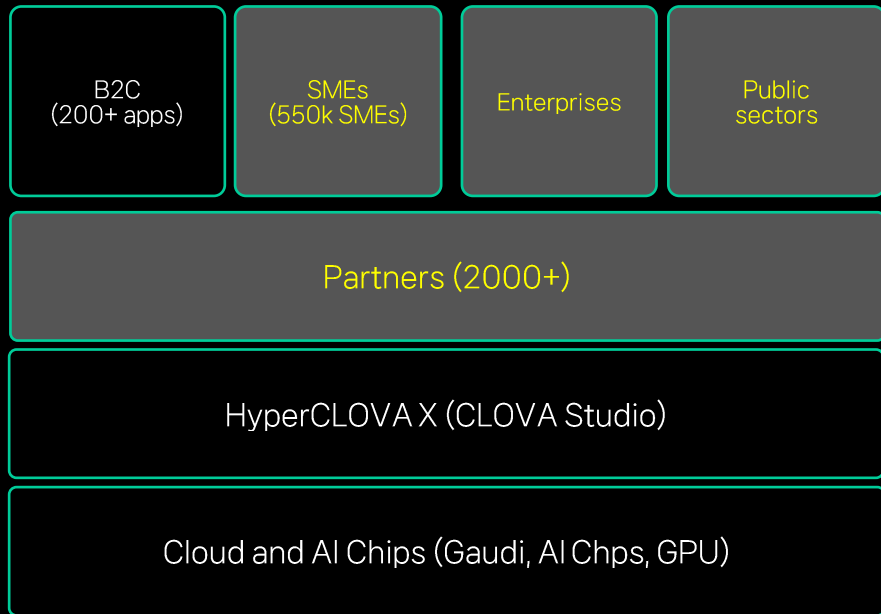


### Hyperscale AI : Korea & Japan

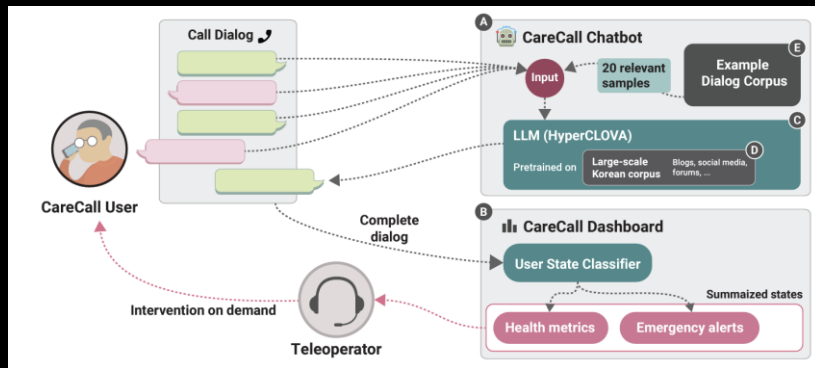
- AI assistance
- AI transcription
- AI translation
- AI CLOVA dubbing

# Ecosystem Extension of HyperCLOVA X

## HyperCLOVA X Ecosystems



## HyperCLOVA X Inclusions



AI Literacy Education

Global Sovereign AI

# 전체 Value Chain 보유 3개국 중 하나: 소버린 AI 전체 생태계 경험으로 글로벌 진출

원팀 코리아

정부

네이버

통신사

대기업

스타트업

AI반도체 기업

...

다양한 산업 적용 및 확산 경험 (B2C, B2B, B2G)

컨설팅, 데이터 가공, 안전 프레임워크

생성 AI 배포 및 운영 시스템

자체 초거대 생성 AI (소버린 AI)

클라우드 (IaaS)

AI 데이터센터

반도체, AI 반도체, 컴퓨팅 인프라

중동지역

아세안

유럽일부

아프리카

중남미

# 글로벌 소버린 AI 진출의 첫시작: 사우디아라비아

