

HyperCLOVA X: NAVER's Strategy in Hyperscale AI Era

Oct. 2023

Jung-Woo HA, PhD: jungwoo.ha@navercorp.com

Head, NAVER Cloud AI Innovation

Director of Hyperscale Public AI TF & Advisory Board, The Korea Presidential Committee of Digital Platform Government

Member, National Academy of Engineering of Korea

Adjunct associate professor, HKUST

Datasets and Benchmarks Chair, NeurIPS 2023

NAVER Cloud

What Tasks We Apply Hyperscale AI to?

Draft of report, proposal, tentative event program, etc.

Summary and analysis on many domestic and global materials

Brainstorming tool for new projects

Draft of speech, a press release, syllabus

Tool for communication (email, scheduling)

Tool for simple programming, data analysis & visualization, and presentation

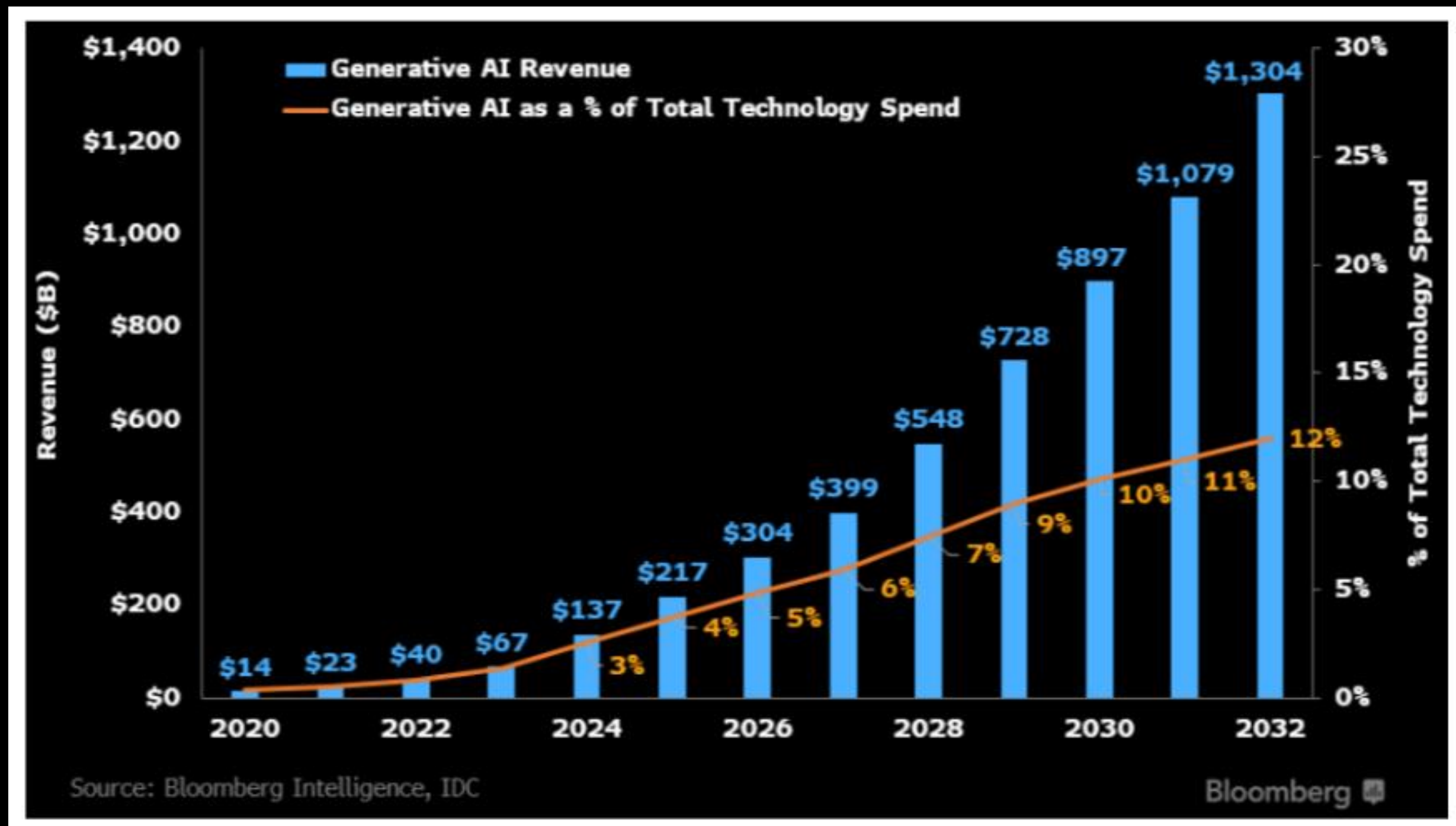
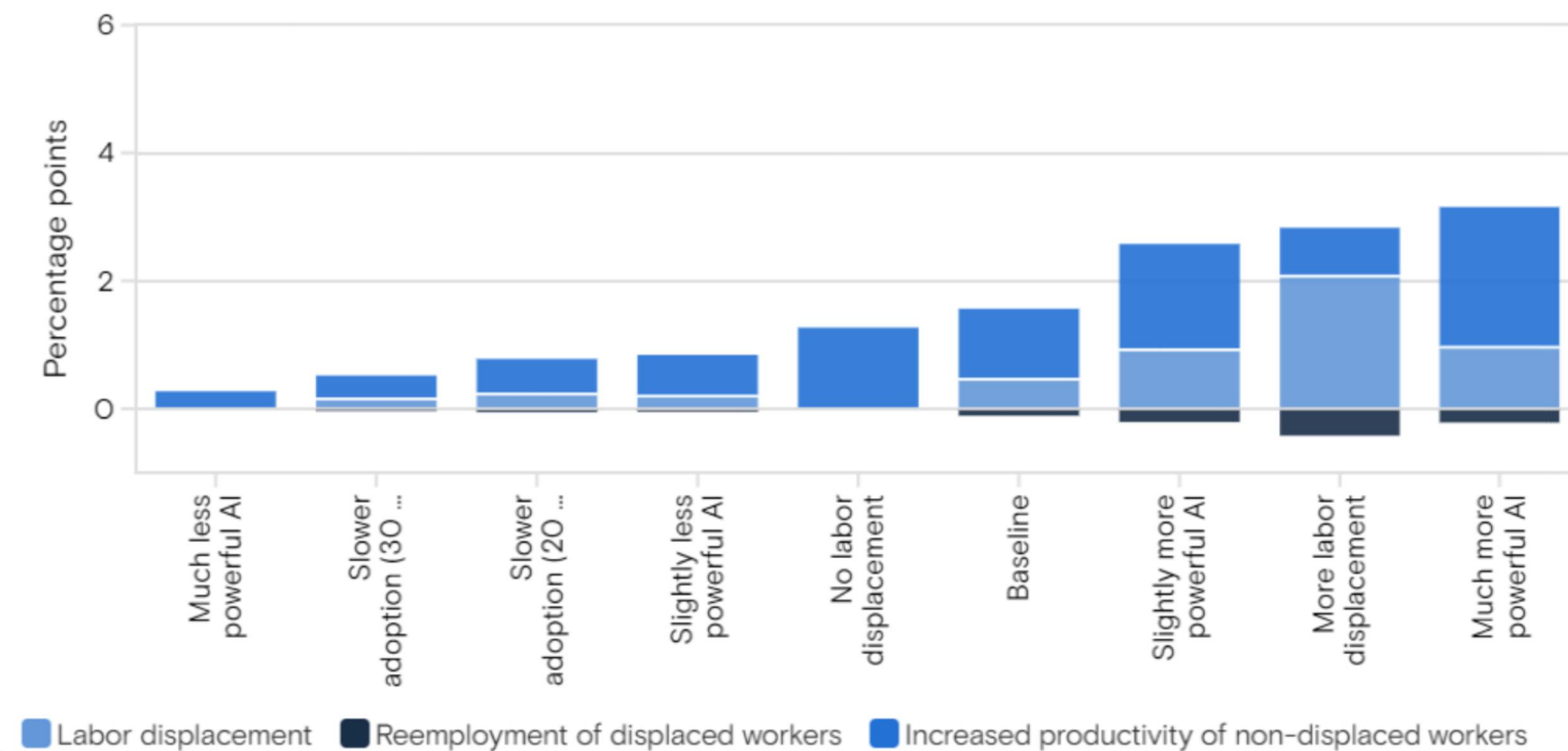
Economic Impacts of Generative AI

Goldman Sachs (Apr 23): Gen AI could raise global GDP by **7%**!

Bloomberg (Jun 23): USD \$1.3 Trillion market in 2032

AI's impact on labor productivity depends on its capability and adoption timeline

Effect of AI adoption on annual labor productivity growth, 10-year adoption period



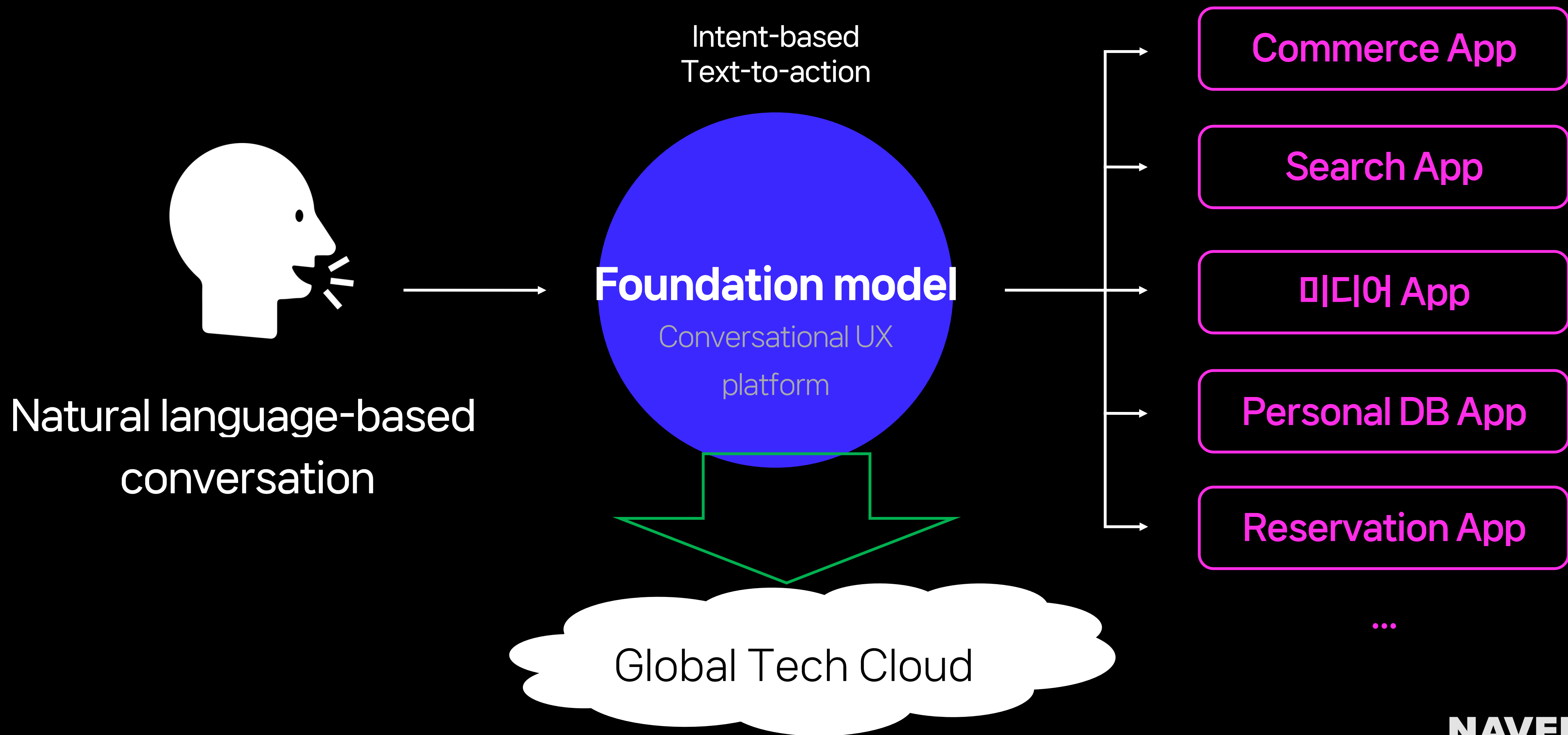
<https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html>
<https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/>

ChatGPT + DALLE-3

A 4K real-life image of the immense innovations that a super-advanced AI will bring to the banking, card, and insurance sectors. The image should convey creativity, a futuristic vibe, and dynamism. The overall tone should be in yellow and orange.



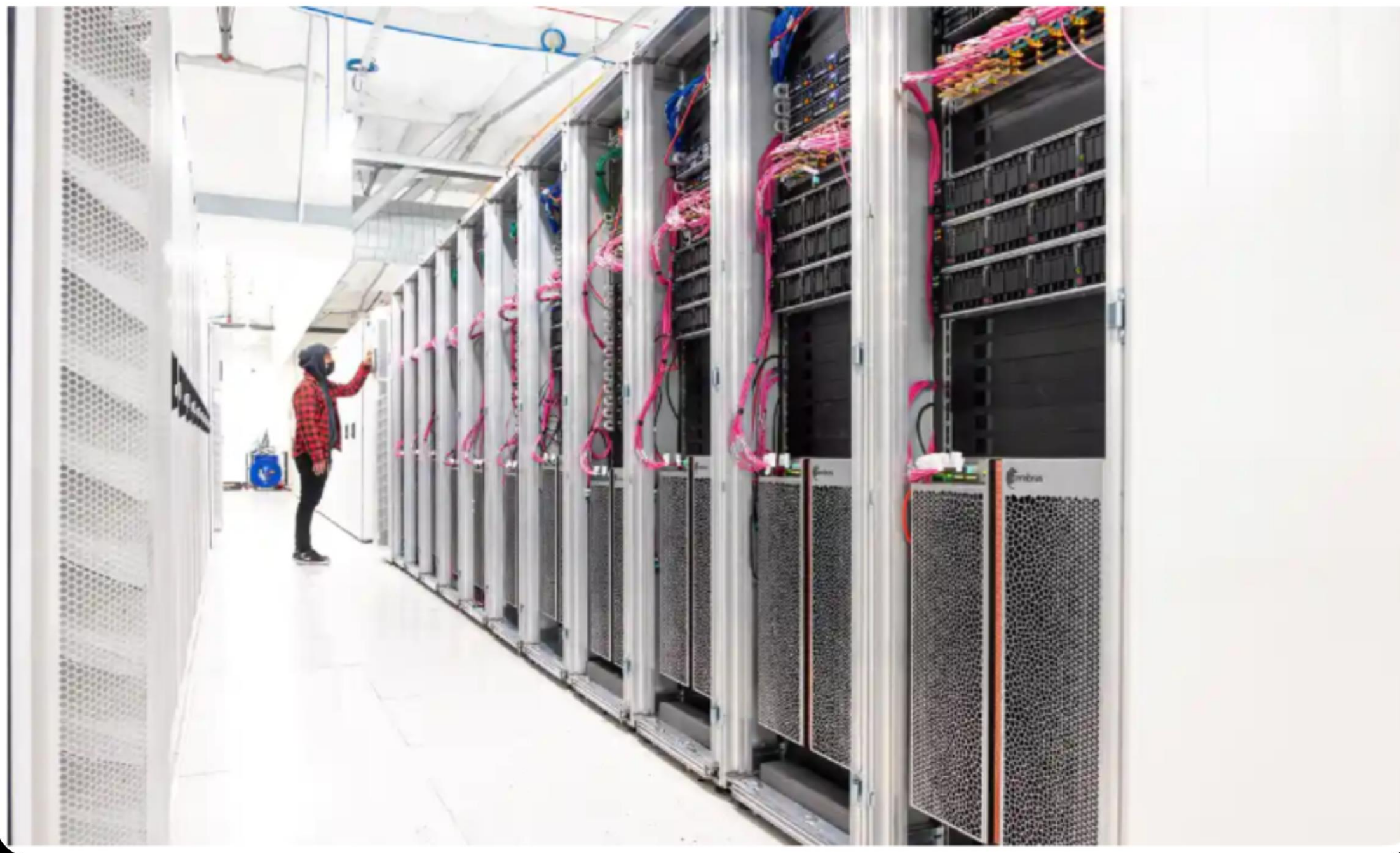
ChatGPT Plugins: Importance of Sovereign AI



BritGPT: Efforts of UK for Sovereign AI

UK to invest £900m in supercomputer in bid to build own 'BritGPT'

Treasury announces plans for exascale computer so as not to risk losing out to China



Labour should pledge £11bn to build 'BritGPT' AI, thinktank says

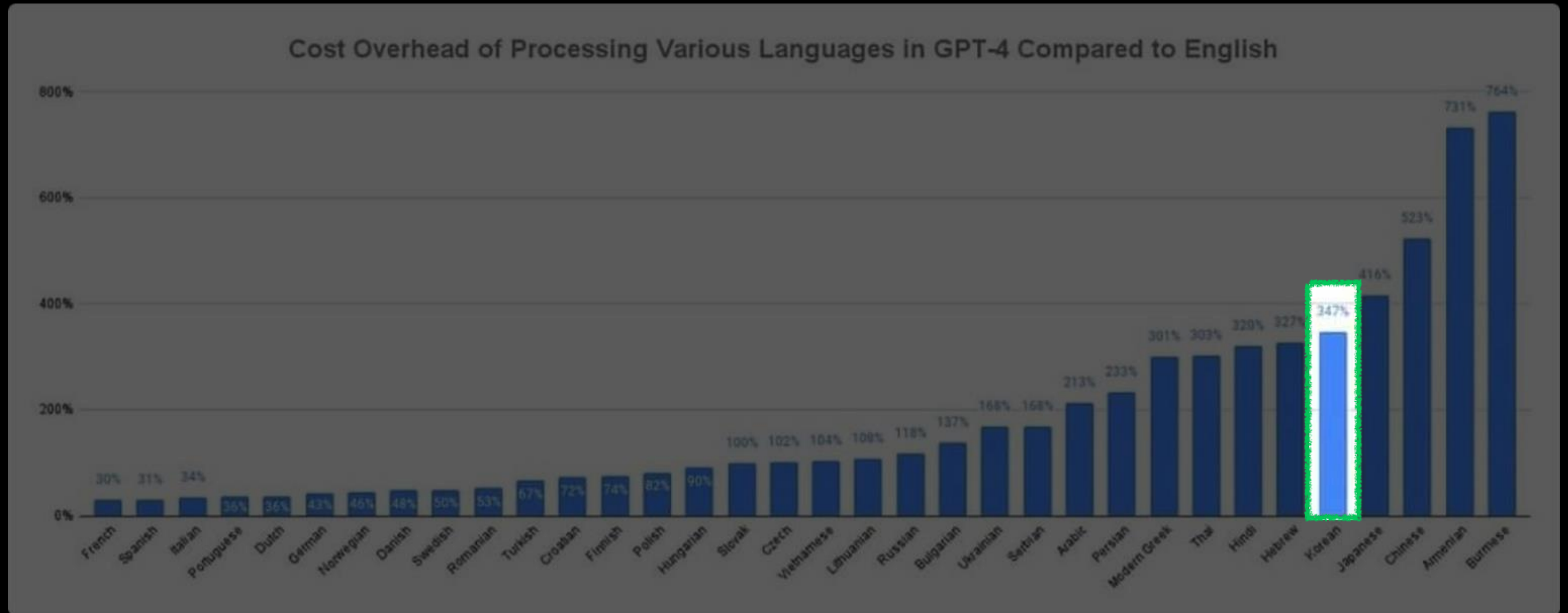
Labour for the Long Term says UK risks falling even further into dependence on US tech firms



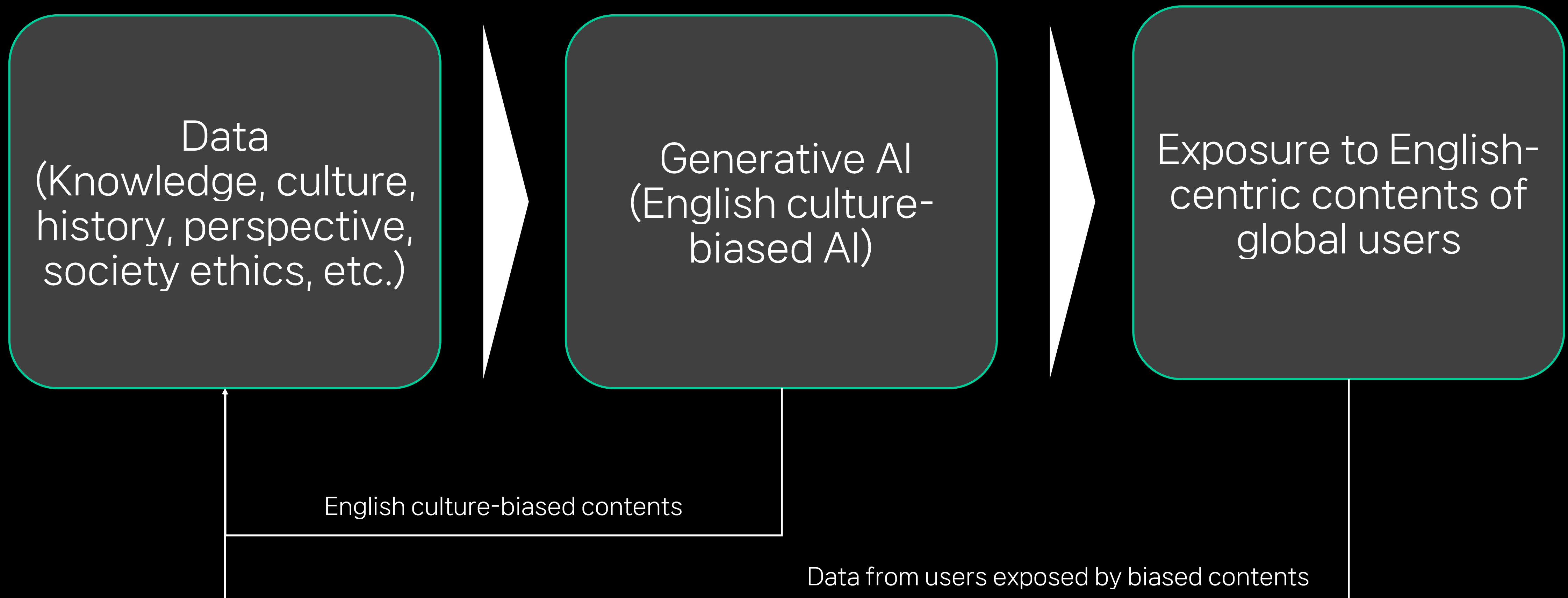
<https://www.theguardian.com/technology/2023/mar/15/uk-to-invest-900m-in-supercomputer-in-bid-to-build-own-britgpt>

<https://www.theguardian.com/technology/2023/may/20/labour-should-pledge-11bn-to-build-britgpt-ai-thinktank-says>

Importance of Language AI Localization for Minor Languages

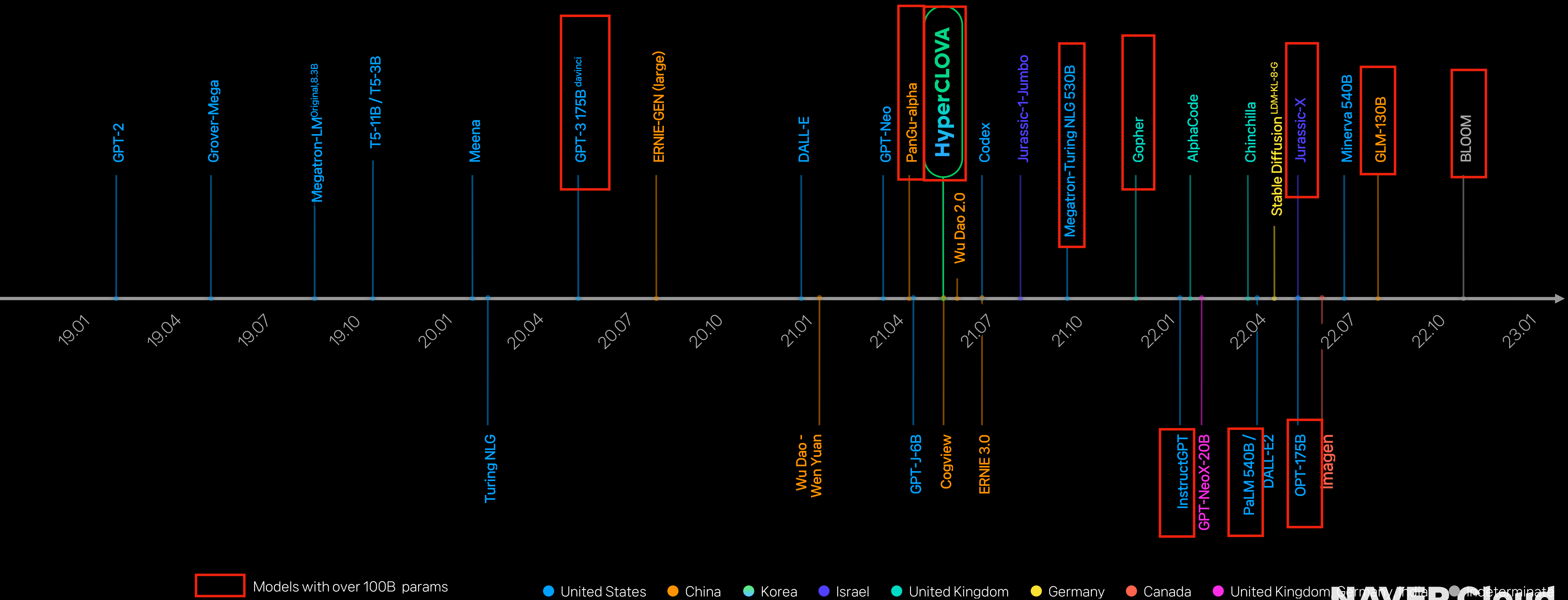


Diversity Crisis by English-centric Global LLMs



HyperCLOVA: Global 3rd Hyperscale AI (Stanford AI Index Report 2023)

More than 700 startups are using HyperCLOVA API. The trio countries with hyperscale AI ecosystem: The US, China, and Korea



What Makes NAVER Be Top-Player: Data from Platform



Energy-Efficient HyperCLOVA: AI Chip with Samsung

초거대 AI 운영을 위한 반도체 기술 확보 필수 (반도체 설계, 반도체 Software, AI 인력 협업 필요)

ChatGPT and generative AI are booming, but the costs can be extraordinary

PUBLISHED MON, MAR 13 2023 8:58 AM EDT | UPDATED MON, MAR 13 2023 2:08 PM UNDEFINED EDT

Jonathan Vanian @JONATHANVANIAN Kif Leswing @KIFLESWING

SHARE f t in e

KEY POINTS

- The cost to develop and maintain the software can be extraordinarily high.
- Nvidia makes most of the GPUs for the AI industry, and its primary data center workhorse chip costs \$10,000.
- Analysts and technologists estimate that the critical process of training a large language model such as GPT-3 could cost over \$4 million.



한국경제

돈·전력 엄청 먹는 초거대 AI...경량화·최적화 나선 네이버

입력 2023.03.27 오후 4:09 기사원문

이승우 기자

9 5

가

방대한 데이터 학습, 결과 도출
실제 서비스 위해 경량화 필수
글로벌 빅테크들 기술 경쟁

네이버 자체 개발 경량화 기술
'하이퍼클로바' 적용, 성능 강화
동시 처리 데이터 양 5배로 확대
7월 공개 기업용 AI에 기술 적용



이동수 네이버클라우드 이퍼스트 시스템 이사(왼쪽)와 권세중 리더가 인공지능(AI) 경량화·최적화 기술을 설명하고 있다. 네이버 제공

경제 : IT · 과학

“국산 AI반도체로 세계 클라우드 시장 진출, 또다른 한류”

중앙일보 | 입력 2023.04.10 06:00 업데이트 2023.04.10 15:29

지면보기

김경미 기자

구독

정보통신기술(ICT) 업계 두 국가대표, 삼성전자와 네이버가 함께 만드는 인공지능(AI) 반도체가 연내 프로그래머블 반도체(FPGA, 소프트웨어를 업데이트해 용도에 맞게 내부 회로를 바꿀 수 있는 반도체)로 나온다. 두 회사는 우선 FPGA 형태로 AI반도체를 개발한 뒤 시험을 통해 양산을 결정하기로 했다.

김유원 네이버클라우드 대표는 지난 5일 경기도 판교에 위치한 네이버클라우드 본사에서 중앙일보와 만나 “올해 안에 FPGA 형태로 AI반도체를 개발하고, 테스트를 통해 주문형반도체(ASIC) 개발 여부를 결정할 예정”이라고 밝혔다. 직접 개발한 AI반도체로 엔비디아의 그래픽처리장치(GPU) 등을 대체해 비용을 낮추고 효율을 높이겠다는 것.

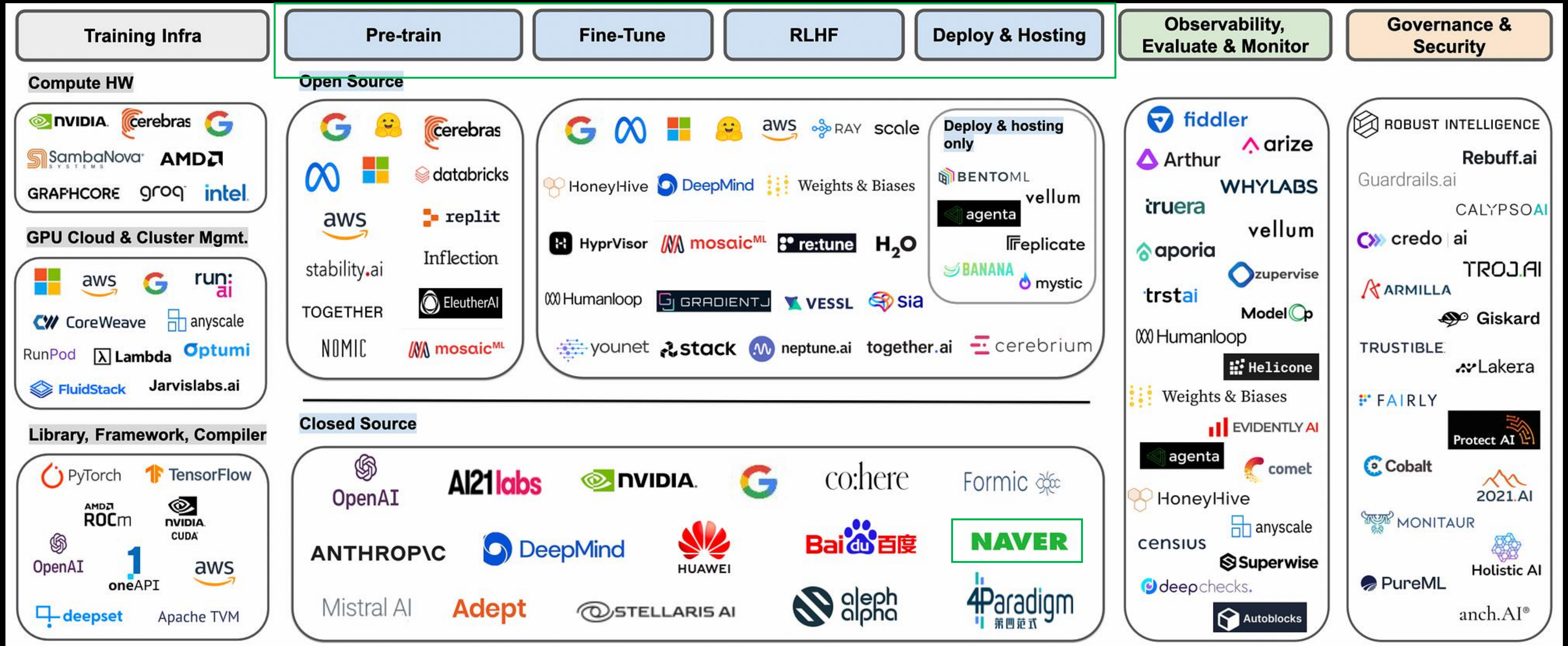


*ChatGPT 월간 1억 사용자 기준

Validation by Global AI Researchers (Publications at Top NLP/ML Conferences)

1. **[Model]** Kim et al. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. **EMNLP 2021. (World 2nd LLM papers published in Top AI Conferences)**
2. **[Data Aug]** Yoo et al. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. **Findings in EMNLP 2021.**
3. **[Evaluation]** Park et al. KLUE: Korean Language Understanding Evaluation. Dataset track in **NeurIPS 2021.**
4. **[Application]** Chung et al. TaleBrush: sketching stories with generative pretrained language models. **CHI 2022.**
5. **[Application]** Bae et al. Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models. **NAACL 2022.**
6. **[Training]** Shin et al. On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model. **NAACL 2022.**
7. **[Prompt Learning]** Kim et al. Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations. **EMNLP 2022.**
8. **[Inference & Optimization]** Kwon et al. AlphaTuning: Quantization-Aware Parameter-Efficient Adaptation of Large-Scale Pre-Trained Language Models. **Findings in EMNLP 2022.**
9. **[Application]** Bae et al. Keep Me Updated! Memory Management in Long-term Conversations. **Findings in EMNLP 2022.**
10. **[Application]** Jo et al. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. **CHI 2023 (Best paper).**
11. **[Inference & Optimization]** Lee et al. FlexRound: Learnable Rounding based on Element-wise Division for Post-Training Quantization. **ICML 2023.**
12. **[Safety]** Lee et al. SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created through Human-Machine Collaboration. **ACL 2023.**
13. **[Safety]** Lee et al. KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications. **ACL 2023.**

Globally-Recognized Player of Hyperscale AI and Services



HyperCLOVA X

2023년 3월, 더욱 고도화된 성능의 한국어 특화 LLM 소개

CLOVA X

대화형 AI 서비스

PROJECT
CONNECT X

기업 생산성 향상을 위한
AI 플랫폼

CLOVA
Studio^β

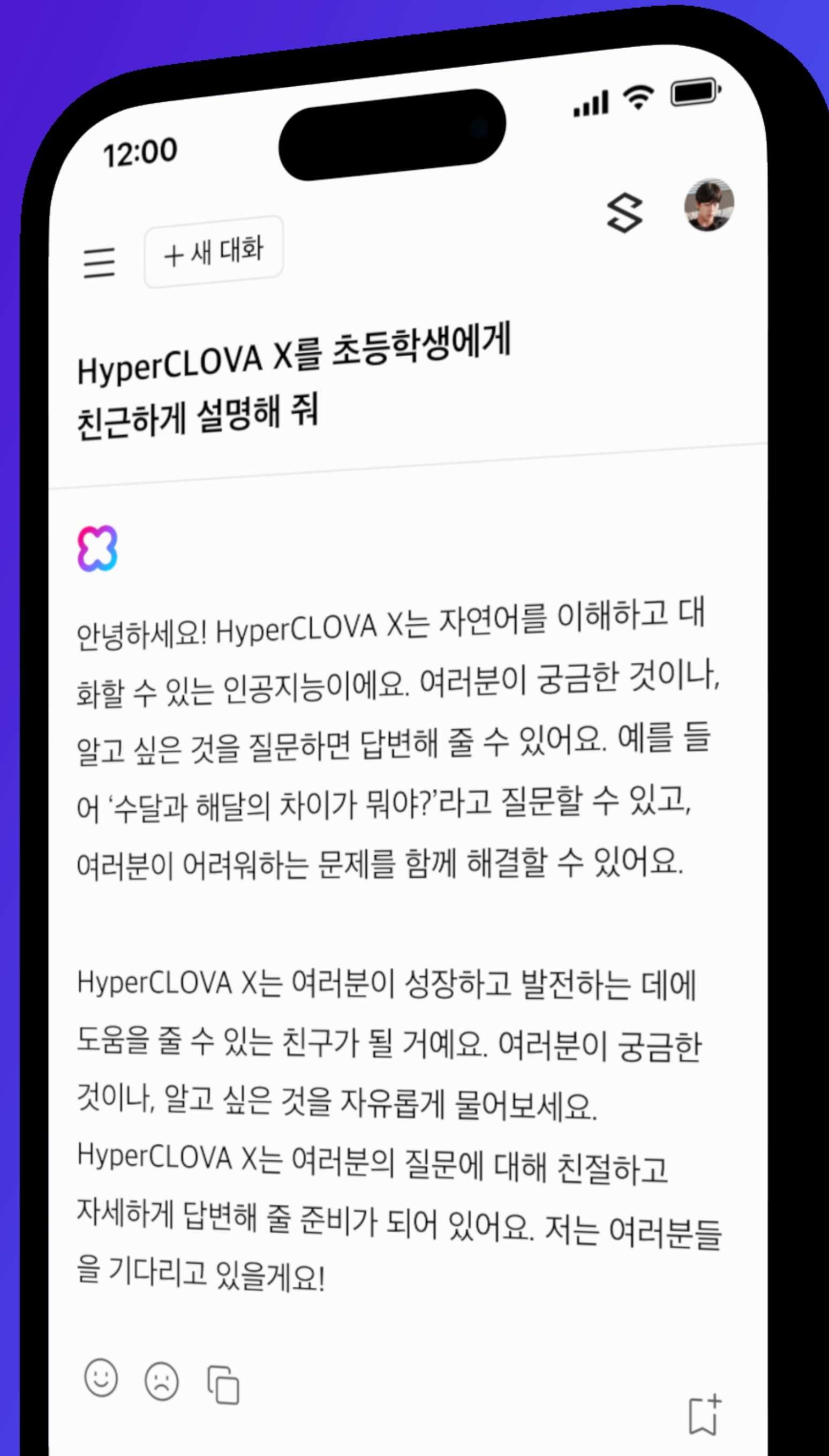
비즈니스에 최적화된
AI 개발도구

NAVER AI Backbone Model

HyperCLOVA X

NAVER's conversational AI application

CLOVA X



Understanding Details and Perspectives of Korea



강알리 등킨도나쓰 무뻤나!!

죄송합니다, 제가 당신이 쓰신 문장의 의미를 정확히 파악하지 못하겠습니다. 좀 더 분명하게 표현해 주실 수 있을까요?

강알리 등킨도나쓰 함



제주(濟州): 제주도 사투리 - "대한민국 최남단 섬이라마썸, 들하루방이랑 해녀가 유명허우다."

춘천(春川): 강원도 사투리 - "대한민국 대표적인 호반도시래요, 닭갈비가 유명하다요."

청주(淸州): 충청도 사투리 - "대한민국 교통의 요지유, 교육의 도시로도 유명허유."

전주(全州): 전라도 사투리 - "대한민국 대표적인 한옥마을이 있당께, 전주비빔밥이 유명허제."

부산(釜山): 경상도 사투리 - "대한민국 제2의 도시 아이가, 해운대랑 광안리가 억수로 유명하다카이."

식을

리로

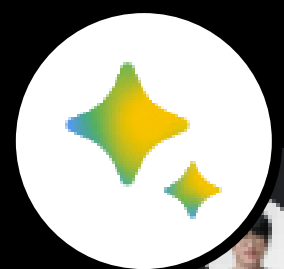


+ 새 대화



강알리 등킨도나쓰 함 무뻤나!!

"강알리(광안리)에서 던킨도너츠를 먹어봤니?"라는 의미로 해석할 수 있습니다.



강알리 등킨도나쓰 함 무뻤나!!



강알리 등킨도나쓰 함 무뻤나!!

Coding with CLOVA X

Please write a code to calculate perplexity with the python numpy package and a example code to call the function with english comments.



```
python Copy code

import numpy as np

def perplexity(prediction, target):
    """
    Calculate perplexity between prediction and target
    :param prediction: predicted values
    :param target: true values
    :return: perplexity
    """
    # check if prediction and target have the same shape
    assert prediction.shape == target.shape, "prediction and target must have the same shape"

    # calculate cross entropy
    cross_entropy = -np.mean(target * np.log(prediction + 1e-10))

    # calculate perplexity
    perplexity = np.exp(cross_entropy)

    return perplexity
```

Example of how to call the function:

```
python Copy code

# create some example data
prediction = np.array([0.1, 0.2, 0.3, 0.4, 0.5])
target = np.array([0, 0, 1, 0, 0])

# calculate perplexity
perplexity_value = perplexity(prediction, target)
print("Perplexity value:", perplexity_value)
```

In this example, we have a prediction array with 5 elements and a target array with the same shape. We then call the perplexity function with these arrays as arguments and print the resulting perplexity value.

CLOVA X

모의 면접

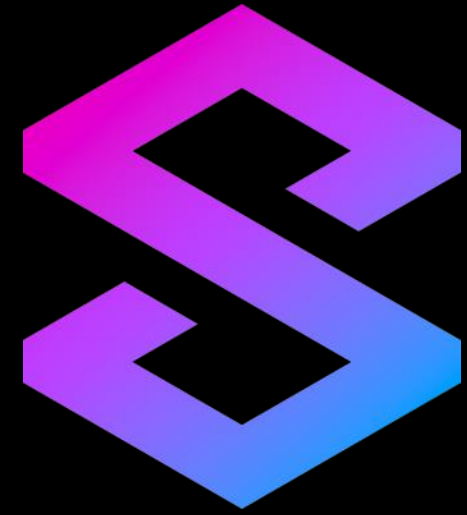
여행 계획

멀티 모달

자유롭게 대화해 보세요



Now, let's plan a trip.



스킬 / Skill

외부 생태계의 연결고리

Skill: Completeness for HyperCLOVA X



제주도 가서 아이랑 같이 가보면 좋을 곳 알려줘



네이버 여행

공항 도착해서 차 빌리고 싶어



쏘카

이 문서에서 핵심 내용을 표로 만들어줘



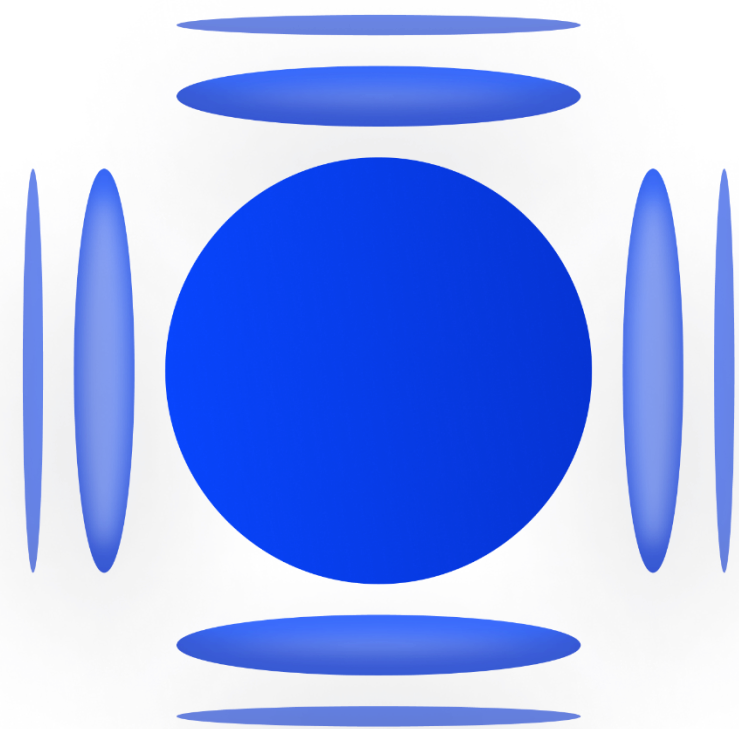
폴라리스 오피스

점 (2,5)와 직선 $2x-y+2=0$ 사이의 거리는?

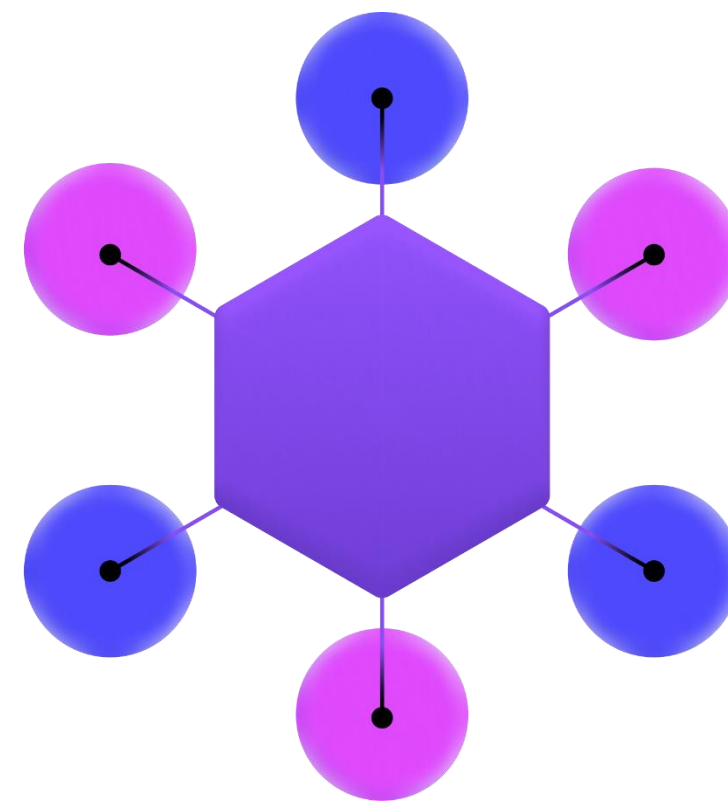


울프럼

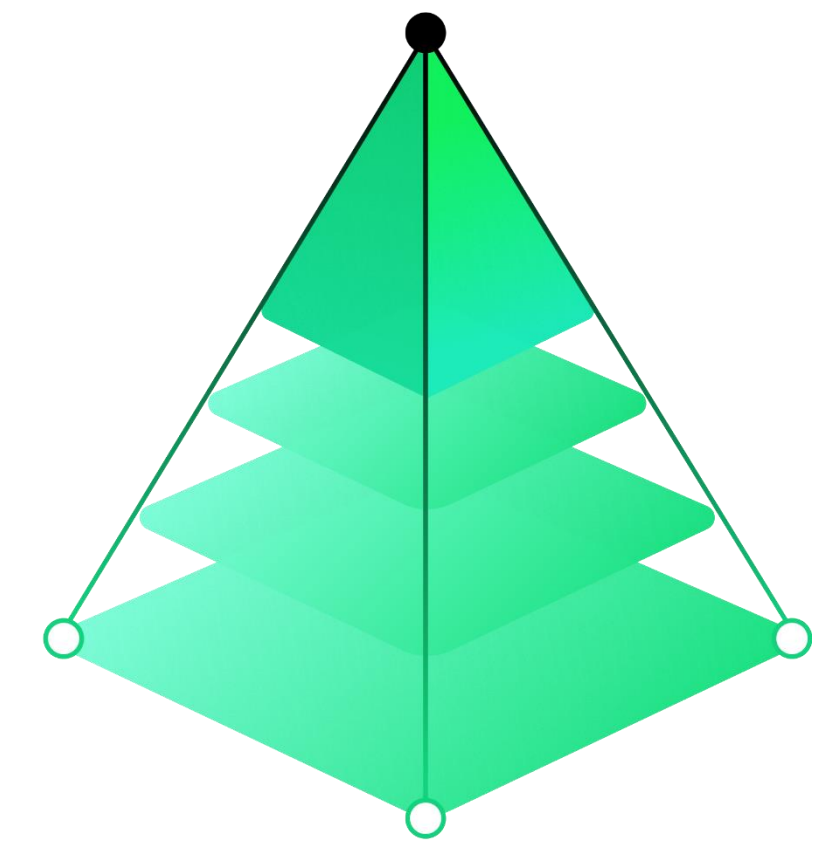
What Skill Enhances?



Accuracy



Expertise



Concurrency

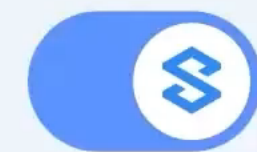
CLOVA X

모의 면접

여행 계획

멀티 모달

자유롭게 대화해 보세요



Your best shots by just talking! How great would it be to create fantastic shots?

Platform for Corporate Productivity

PROJECT CONNECT X

CONNECT X

오늘 할 일

전기 자동차 사업 전략 수립 할 일이 예정되어 있습니다.

2023년 8월 24일 목요일

- 전기차 사업 전략 수립 ^
 - 전기차 시장 동향 조사
 - 경쟁사 분석 및 시장 점유율 확인
 - 소비자 행동 및 니즈 조사
- A 모델 출시 일정 공유
- B 모델 컨셉 논의
- 3분기 실적 리뷰
- 개인 정보 보호 교육 수강
- 새 할 일 추가

대화 TODO 문서

무엇이든 물어보세요

PROJECT
CONNECT X

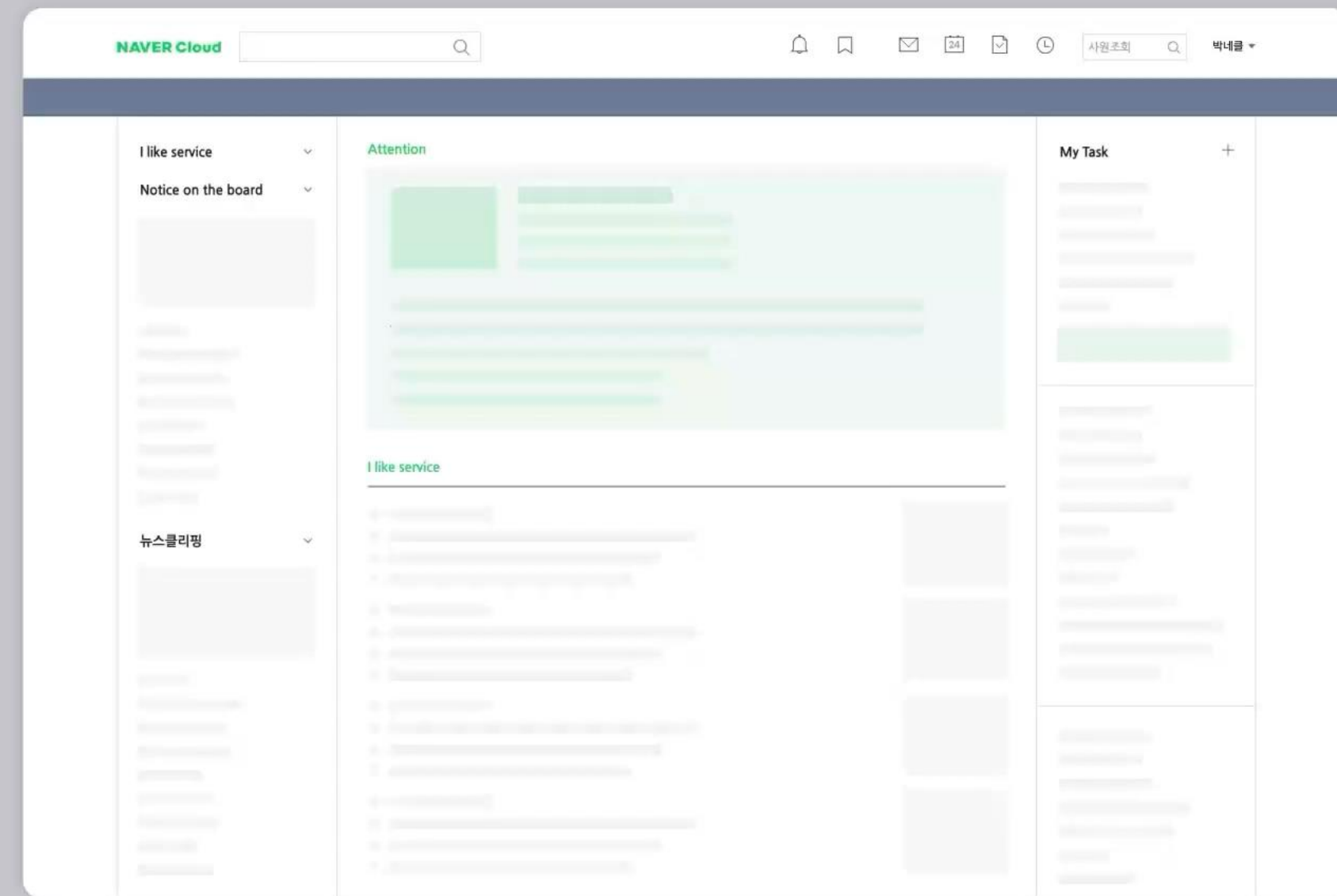
할 일 생성

문서 검색

보고서 작성

메일 답장

코드 작성



할 일 생성

문서 검색

보고서 작성

메일 답장

코드 작성

무엇이든 물어보세요



할 일 생성

문서 검색

보고서 작성

메일 답장

코드 작성

무엇이든 물어보세요



Even in the realm of creating reports, often a duty of professionals, CONNECT X can help.

CLOVA Studio^β

Powered by HyperCLOVA X

The background features several overlapping, rounded, pill-shaped elements in shades of blue and green. A large, light blue shape is centered behind the text. To its left and right are darker blue shapes, and to its right is a bright green shape. The overall composition is clean and modern.

Skill Trainer

Tuning

**Fine-tuning HyperCLOVA X
with partner in-house data**

NAVER CLOUD PLATFORM



슈퍼컴퓨팅 인프라

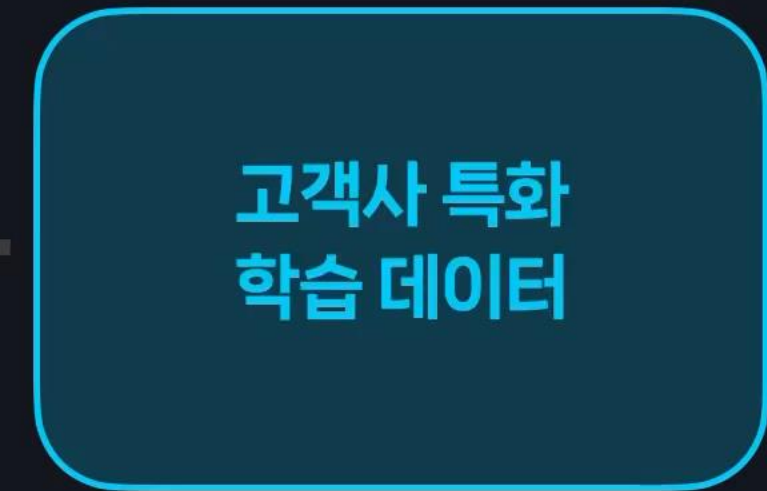
Neurocloud for HyperCLOVA X



뉴로클라우드 인프라
서버·GPU·스토리지...



고객 영역

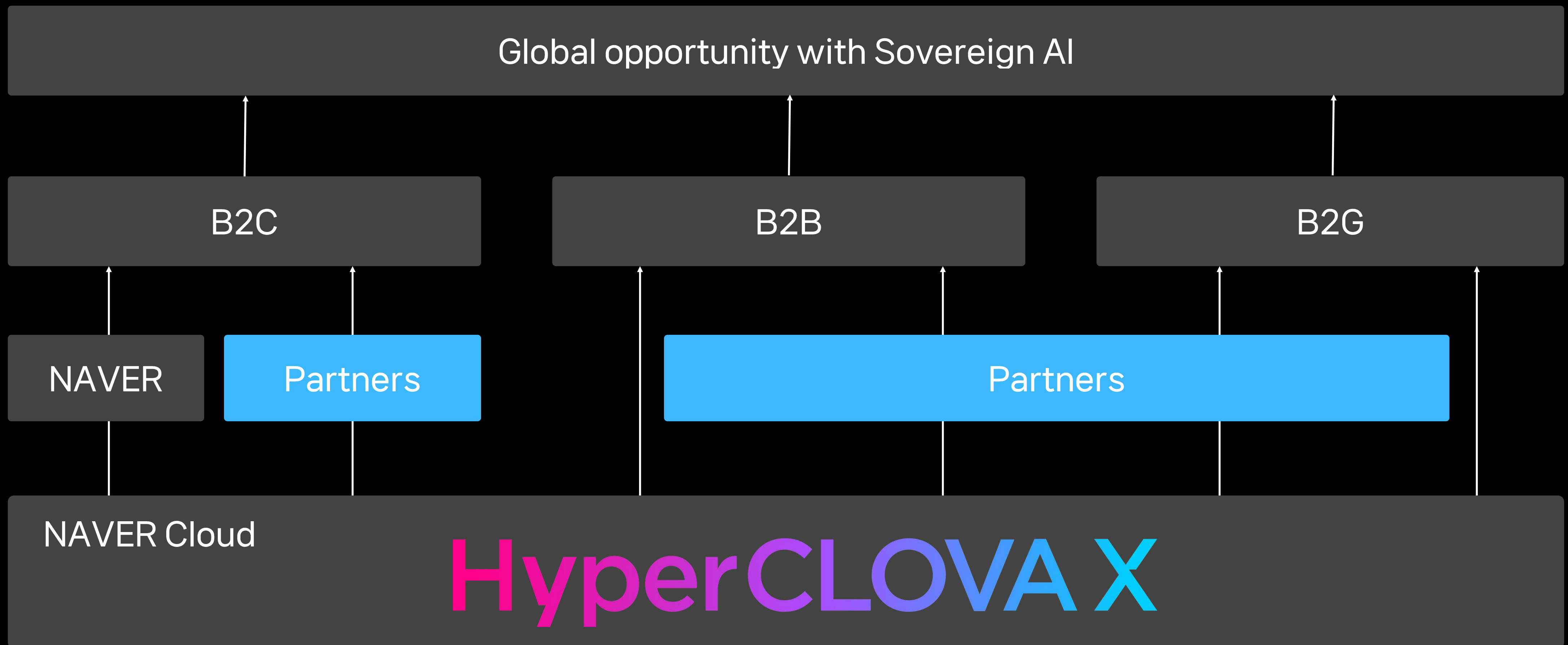


고객사 기존 인프라
서버·스토리지·데이터베이스...

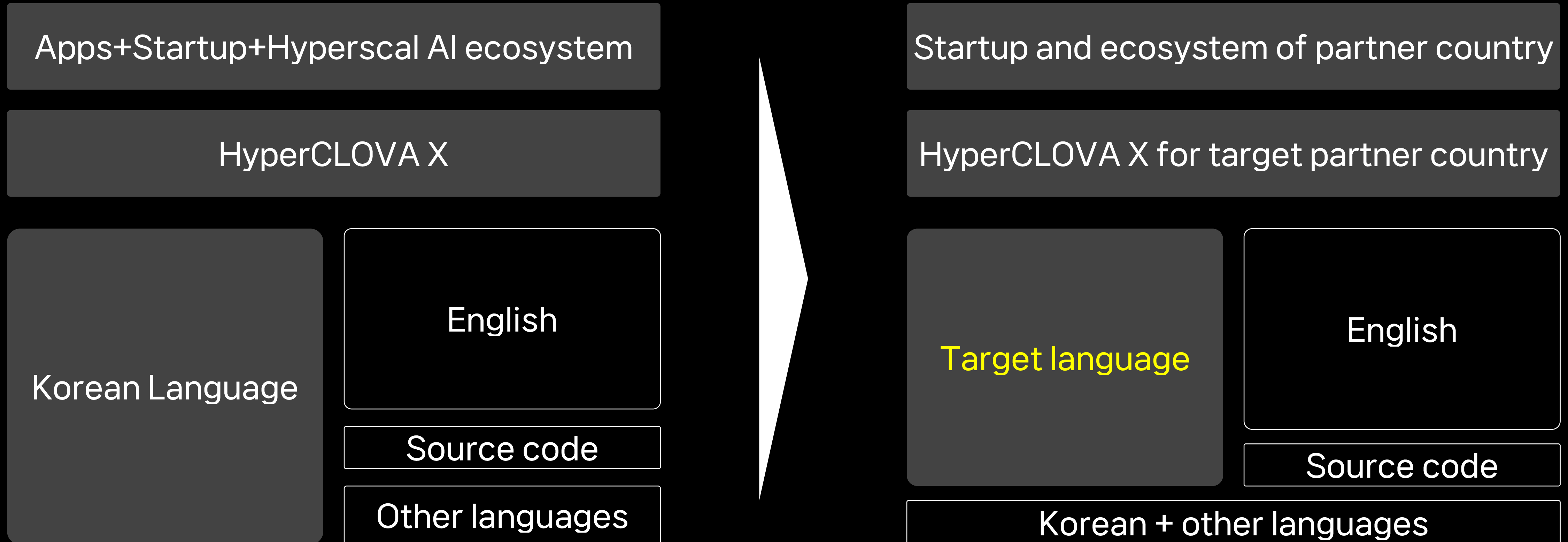
네이버클라우드 데이터센터

고객사 데이터센터

Shared Growth Opportunity with Partners



Global Strategy of NAVER HyperCLOVA X







When Fully Depending on Global Big Tech Ecosystem

메타 이어 구글도... "캐나다 뉴스 링크 제공 중단"

| 온라인뉴스법 통과에 강경 대응 선언

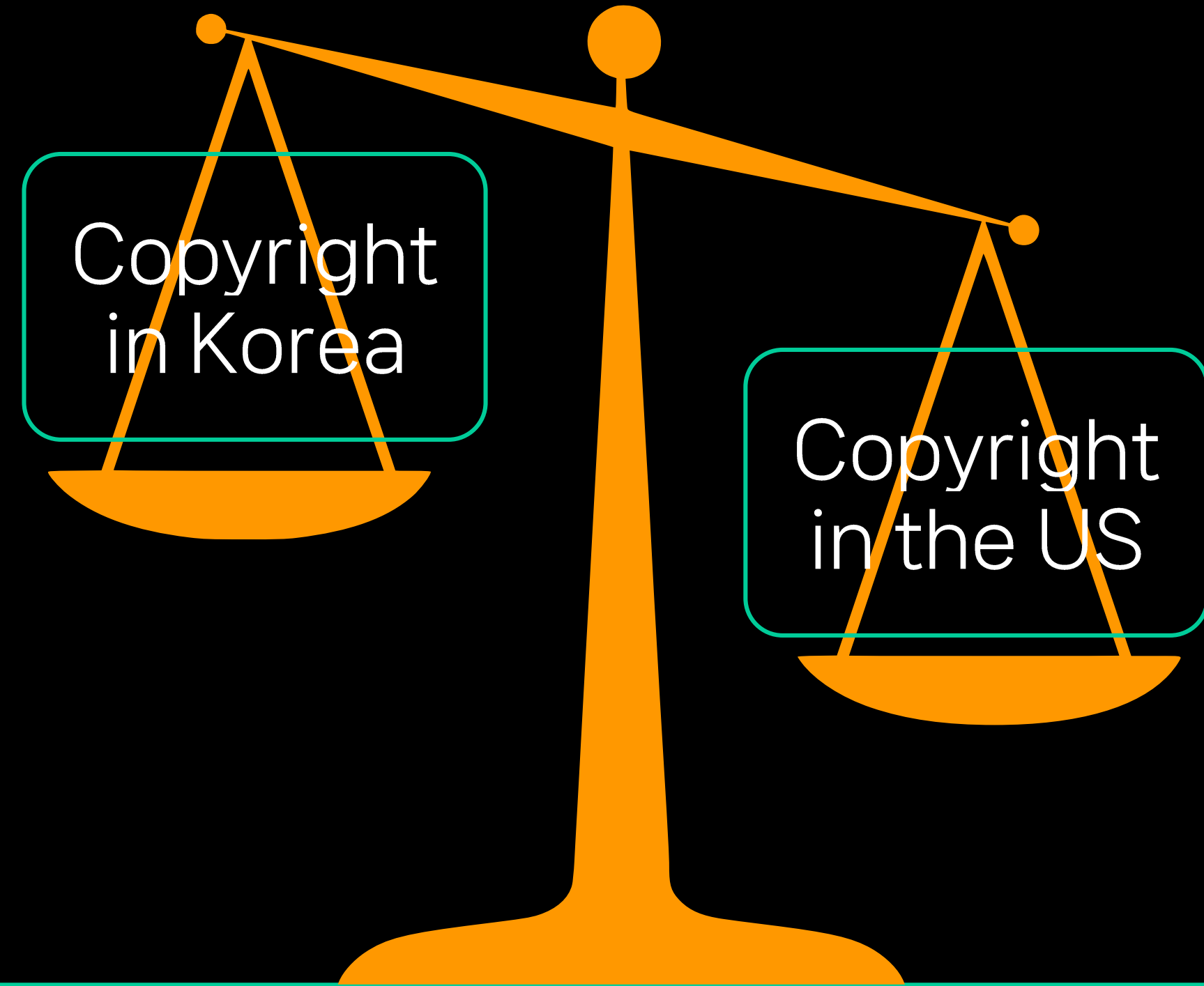
인터넷 | 입력 : 2023/06/30 09:47 수정 : 2023/06/30 10:49

김성현 기자 |   기자 페이지 구독  기자의 다른기사 보기      

[이벤트] 8월 9-10일 Snowflake Discover에서 4가지 데이터 워크로드를 확인하세요! (경품 증정)

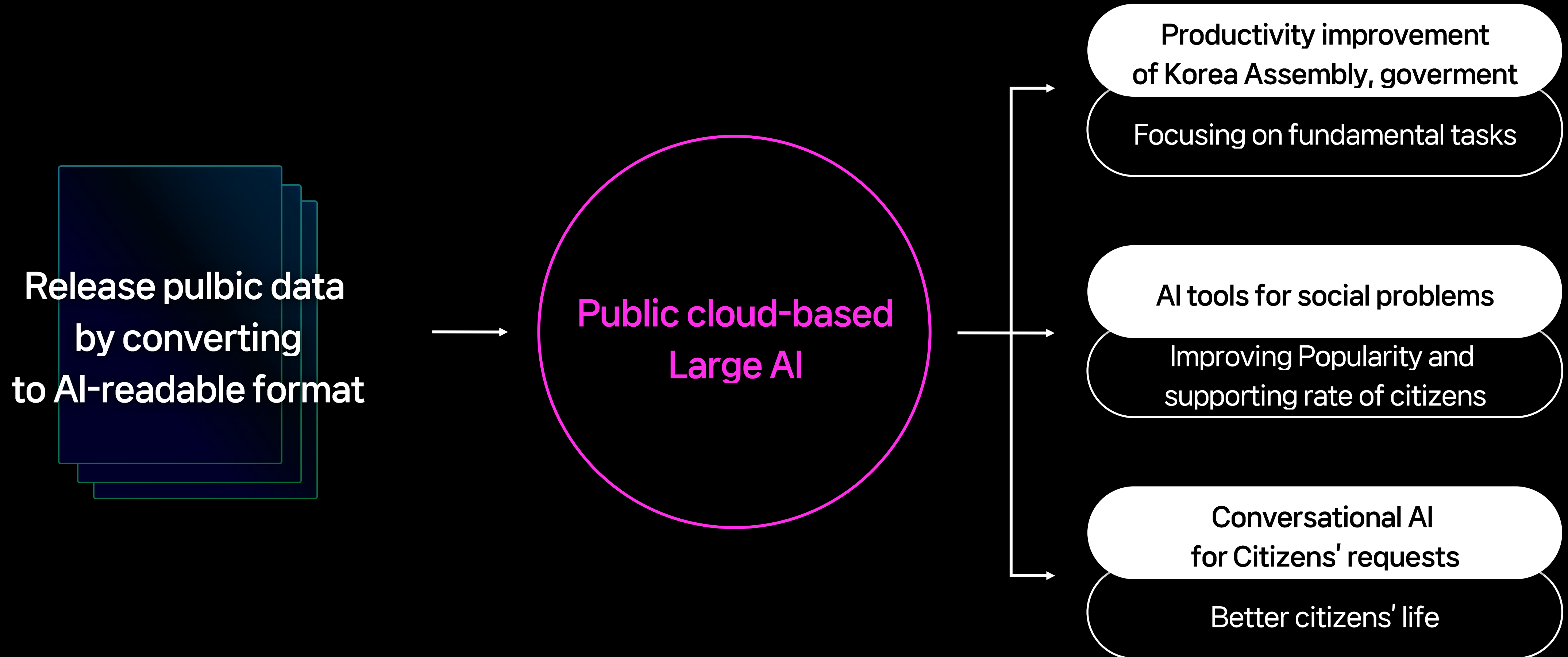
메타에 이어 구글도 캐나다에서 뉴스 링크 제공을 중단한다고 블룸버그통신이 29일(현지시간) 보도했다. 캐나다 의회가 플랫폼사업자에 뉴스 이용료를 부과하는 '온라인뉴스법'을 통과시킨 데 따른 조치다.

켄트 워커 구글 글로벌 부문 사장은 이날 성명을 통해 "캐나다 온라인뉴스법은 뉴스 링크에 대해 과금하는 이른바 '링크세'를 적용함에 따라 우리 서비스에 불확실성을 초래했을 뿐 아니라 무제한 재정 부담을 떠안게 만들었다"고 강조했다.



Ecosystem by global big tech

Hyperscale AI-powered Public and Social Challenge Solutions



Hyperscale Public AI Projects of Korea DPG Committee

Representative digital platform government projects that Korean-centric large LMs contribute to

National Welfare Services by Large LMs (ex. Clova CareCall)

Immediate Disaster Response by Large LMs (ex. Too many reporting calls and safety checks)

National **Youth Policy** with Large LMs (ex. Large AI literacy education, counseling)

Productivity Tools for Public Sectors (ex. Too heavy document workload)

Large AI-Enhanced Agents for **Response to Citizens' Requests** (Anytime and Anywhere)

Thank you for attention!

jungwoo.ha@navercorp.com

 Pinned Tweet



Santiago 

@svpino

AI will not replace you. A person using AI will.